Main

Volume 72B
Number 4
Oct.–Dec. 1968

FEB 10 1969

FEB 10 1969

Mathematical Sciences

Formerly—
Mathematics and
Mathematical Physics

MIAMI UNIVERSITY LIBRARY

U.S. DEPARTMENT
OF COMMERCE

SECTION B

JOURNAL OF RESEARCH

NATIONAL BUREAU OF STANDARDS

Vol. 72B, No. 4 • Journal of Research of the National Bureau of Standards • October–December 1968

Journal of Research

of the

National Bureau of Standards

B. Mathematical Sciences
(Formerly Mathematics and Mathematical Physics)

OCT.–DEC. • 1968

VOLUME • 72B

NUMBER • 4

Editor: Morris Newman

Associate Editors: John J. Sopka, F. W. Olver, and J. R. Edmonds

Publication dates: Mar. 10, June 10, Sept. 10, Dec. 10, 1968

JOURNAL OF RESEARCH

The National Bureau of Standards Journal of Research reports research and development in the fields of activity shown at right. Also included from time to time are survey articles on topics closely related to the Bureau's scientific and technical program.

The Journal is published in three separate sections as follows:

A. Physics and Chemistry

Contains papers of interest primarily to scientists working in these fields. Issued six times a year. Annual subscription: domestic, $6.00; foreign, $7.25. Single copy $1.00.

B. Mathematical Sciences

Presents studies and compilations designed mainly for the mathematician and the theoretical physicist. Issued quarterly. Annual subscription: domestic, $2.25; foreign, $2.75. Single copy 75c.

C. Engineering and Instrumentation

Reports research and development results of interest chiefly to the engineer and the applied scientist. Issued quarterly. Annual subscription: domestic, $2.75; foreign, $3.50. Single copy 75c.

U.S. DEPARTMENT OF COMMERCE
C. R. Smith, Secretary
NATIONAL BUREAU OF STANDARDS
A. V. Astin, Director

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards[1] was established by an act of Congress March 3, 1901. Today, in addition to serving as the Nation's central measurement laboratory, the Bureau is a principal focal point in the Federal Government for assuring maximum application of the physical and engineering sciences to the advancement of technology in industry and commerce. To this end the Bureau conducts research and provides central national services in three broad program areas and provides central national services in a fourth. These are: (1) basic measurements and standards, (2) materials measurements and standards, (3) technological measurements and standards, and (4) transfer of technology.

The Bureau comprises the Institute for Basic Standards, the Institute for Materials Research, the Institute for Applied Technology, and the Center for Radiation Research.

THE INSTITUTE FOR BASIC STANDARDS provides the central basis within the United States of a complete and consistent system of physical measurement, coordinates that system with the measurement systems of other nations, and furnishes essential services leading to accurate and uniform physical measurements throughout the Nation's scientific community, industry, and commerce. The Institute consists of an Office of Standard Reference Data and a group of divisions organized by the following areas of science and engineering:

Applied Mathematics—Electricity—Metrology—Mechanics—Heat—Atomic Physics—Cryogenics[2]—Radio Physics[2]—Radio Engineering[2]—Astrophysics[2]—Time and Frequency.[3]

THE INSTITUTE FOR MATERIALS RESEARCH conducts materials research leading to methods, standards of measurement, and data needed by industry, commerce, educational institutions, and government. The Institute also provides advisory and research services to other government agencies. The Institute consists of an Office of Standard Reference Materials and a group of divisions organized by the following areas of materials research:

Analytical Chemistry—Polymers—Metallurgy—Inorganic Materials—Physical Chemistry.

THE INSTITUTE FOR APPLIED TECHNOLOGY provides for the creation of appropriate opportunities for the use and application of technology within the Federal Government and within the civilian sector of American industry. The primary functions of the Institute may be broadly classified as programs relating to technological measurements and standards and techniques for the transfer of technology. The Institute consists of a Clearinghouse for Scientific and Technical Information,[3] a Center for Computer Sciences and Technology, and a group of technical divisions and offices organized by the following fields of technology:

Building Research—Electronic Instrumentation—Technical Analysis—Product Evaluation—Invention and Innovation—Weights and Measures—Engineering Standards—Vehicle Systems Research.

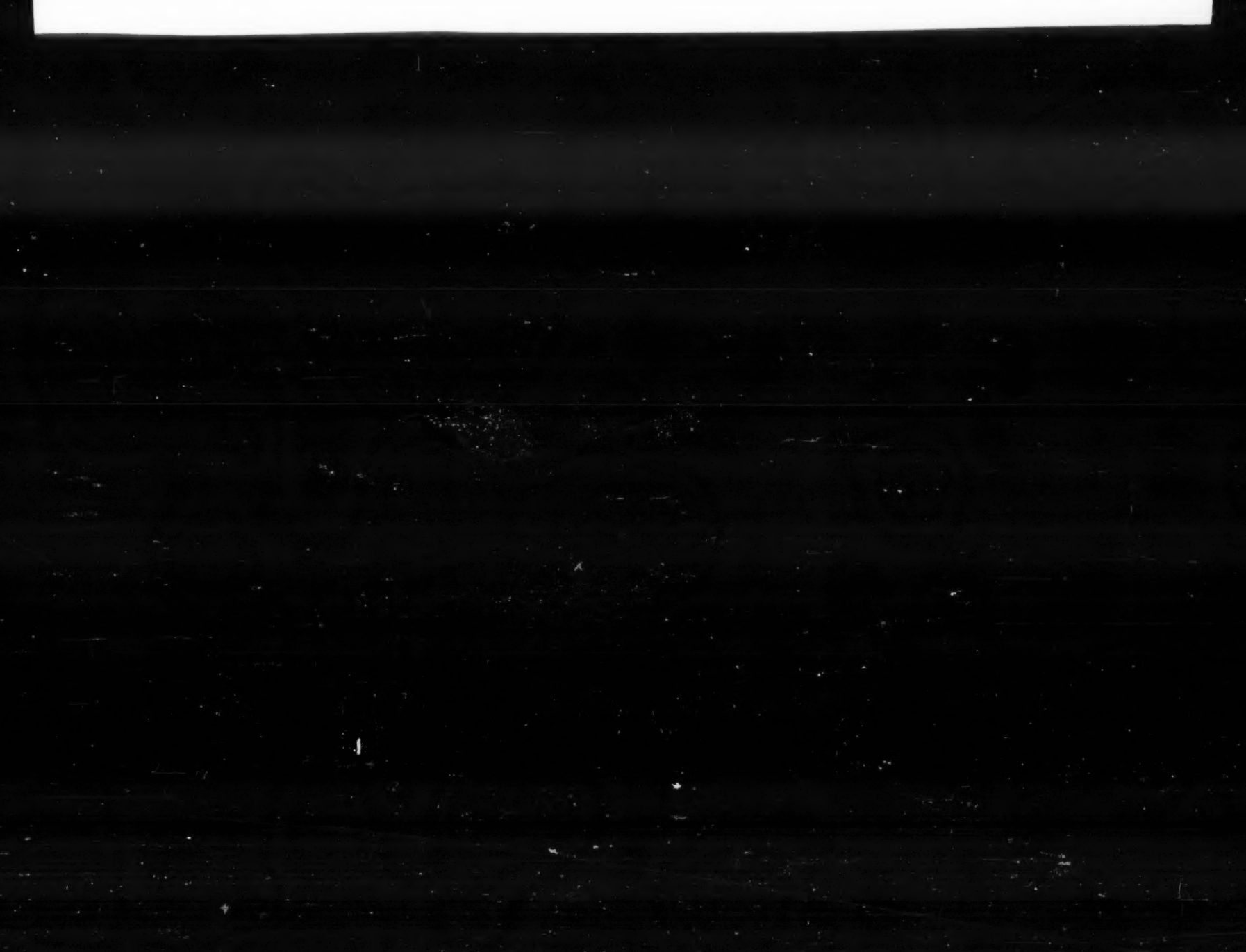
THE CENTER FOR RADIATION RESEARCH engages in research, measurement, and application of radiation to the solution of Bureau mission problems and the problems of other agencies and institutions. The Center for Radiation Research consists of the following divisions:

Reactor Radiation—Linac Radiation—Applied Radiation—Nuclear Radiation.

[1] Headquarters and Laboratories at Gaithersburg, Maryland, unless otherwise noted; mailing address Washington, D.C. 20234.
[2] Located at Boulder, Colorado 80302.
[3] Located at 5285 Port Royal Road, Springfield, Virginia 22151.

Order all publications from the Superintendent of Documents
U.S. Government Printing Office, Washington, D.C. 20402.

Use of funds for printing this publication approved by the Director of the Bureau of the Budget (June 22, 1966).



JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 72B, No. 4, October–December 1968

Principal Submatrices VII: Further Results Concerning Matrices With Equal Principal Minors*

R. C. Thompson**

(July 30, 1968)

This paper characterizes real symmetric matrices A such that all $t \times t$ principal minors are equal and all $t \times t$ nonprincipal minors are of fixed sign, for two consecutive values of t less than rank A. It also characterizes matrices A (over an arbitrary field) in which all $t \times t$ principal minors are equal and all nonprincipal $t \times t$ minors are equal, for one fixed value of t less than rank A.

Key Words: Matrix, principal submatrices; rank; symmetric matrix.

In the paper "Principal Submatrices V," [5],[1] a classification was found for symmetric matrices A for which all $t \times t$ principal minors of A are equal, for *three* consecutive values of t less than the rank of A. It is the purpose of this paper to present a similar theorem classifying the real symmetric matrices in which the condition on the principal minors is weakened to requiring that all $t \times t$ principal minors of A be equal, for *two* consecuitive values of t less than the rank of A, and in which a sign condition is imposed on the nonprincipal $t \times t$ minors for these two consecutive values of t. This result is presented in Theorem 1. In this paper we also classify all square matrices A (over an arbitrary field and not necessarily symmetric) in which the condition on the principal minors of A is weakened to requiring that all $t \times t$ principal minors of A be equal for *one* value of t less than the rank of A, and for this value of t the condition on the nonprincipal $t \times t$ minors of A is strengthened to requiring that they all be equal. This result is presented in Theorem 4.

THEOREM 1. *Let* r *be a fixed integer and let* A *be an* n × n *symmetric matrix over the real number field, such that:*

(i) *all* r × r *principal minors of* A *are equal;*

(ii) *all* (r + 1) × (r + 1) *principal minors of* A *are equal;*

(iii) *all nonprincipal* r × r *minors of* A *which do not vanish have a common sign;*

(iv) *all nonprincipal* (r + 1) × (r + 1) *minors of* A *which do not vanish have a common sign;*

(v) *rank* A ≥ r + 2.

Then A *is scalar:* $A = aI_n$.

PROOF. The proof follows the pattern of the analogous theorem (Theorem 13) in [5].

First case. Let $r = 1$. Let $A = (a_{ij})$. The equality of the 1×1 principal minors forces all a_{ii} to be equal, say $a_{ii} = a$. The equality of all 2×2 principal minors forces all a_{ij}^2 to be equal, for $i \neq j$. Thus $a_{ij} = \pm b$, for $i \neq j$. Because all nonvanishing nonprincipal 1×1 minors have a common sign, we see that (choosing b properly) all $a_{ij} = b$ for $i \neq j$. We wish to show that $b = 0$. The nonprincipal 2×2 minor $\det A[1, 2|1, 3] = b(a - b)$ and the nonprincipal 2×2 minor $\det A[2, 3|1, 2] = -b(a - b)$. The sign condition on these nonprincipal 2×2 minors now shows that $b = 0$ or $a - b = 0$. If $a - b = 0$ then $A = bJ_n$, where J_n is the $n \times n$ matrix in which each entry is one. Since J_n has rank one, the possibility $a = b$ contradicts hypothesis (v). Therefore $b = 0$ and hence $A = aI_n$ as claimed.

*An invited paper. The preparation of this paper was supported in part by the U.S. Air Force Office of Scientific Research, under Grant 698-67.

**Present address: Department of Mathematics, University of California, Santa Barbara, Calif. 93106.

[1] Figures in brackets indicate the literature references at the end of this paper.

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 72B, No. 4, October–December 1968

Principal Submatrices VII: Further Results Concerning Matrices With Equal Principal Minors*

R. C. Thompson**

(July 30, 1968)

This paper characterizes real symmetric matrices A such that all $t \times t$ principal minors are equal and all $t \times t$ nonprincipal minors are of fixed sign, for two consecutive values of t less than rank A. It also characterizes matrices A (over an arbitrary field) in which all $t \times t$ principal minors are equal and all nonprincipal $t \times t$ minors are equal, for one fixed value of t less than rank A.

Key Words: Matrix; principal submatrices; rank; symmetric matrix.

In the paper "Principal Submatrices V," [5],[1] a classification was found for symmetric matrices A for which all $t \times t$ principal minors of A are equal, for *three* consecutive values of t less than the rank of A. It is the purpose of this paper to present a similar theorem classifying the real symmetric matrices in which the condition on the principal minors is weakened to requiring that all $t \times t$ principal minors of A be equal, for *two* consecuitive values of t less than the rank of A, and in which a sign condition is imposed on the nonprincipal $t \times t$ minors for these two consecutive values of t. This result is presented in Theorem 1. In this paper we also classify all square matrices A (over an arbitrary field and not necessarily symmetric) in which the condition on the principal minors of A is weakened to requiring that all $t \times t$ principal minors of A be equal for *one* value of t less than the rank of A, and for this value of t the condition on the nonprincipal $t \times t$ minors of A is strengthened to requiring that they all be equal. This result is presented in Theorem 4.

THEOREM 1. *Let* r *be a fixed integer and let* A *be an* n × n *symmetric matrix over the real number field, such that:*

(i) *all* r × r *principal minors of* A *are equal;*

(ii) *all* (r + 1) × (r + 1) *principal minors of* A *are equal;*

(iii) *all nonprincipal* r × r *minors of* A *which do not vanish have a common sign;*

(iv) *all nonprincipal* (r + 1) × (r + 1) *minors of* A *which do not vanish have a common sign;*

(v) $\overline{rank}\ A \geq r+2$.

Then A *is scalar:* $A = aI_n$.

PROOF. The proof follows the pattern of the analogous theorem (Theorem 13) in [5].

First case. Let $r=1$. Let $A=(a_{ij})$. The equality of the 1×1 principal minors forces all a_{ii} to be equal, say $a_{ii}=a$. The equality of all 2×2 principal minors forces all a_{ij}^2 to be equal, for $i \neq j$. Thus $a_{ij}=\pm b$, for $i \neq j$. Because all nonvanishing nonprincipal 1×1 minors have a common sign, we see that (choosing b properly) all $a_{ij}=b$ for $i \neq j$. We wish to show that $b=0$. The nonprincipal 2×2 minor $\det A[1, 2|1, 3]=b(a-b)$ and the nonprincipal 2×2 minor $\det A[2, 3|1, 2]=-b(a-b)$. The sign condition on these nonprincipal 2×2 minors now shows that $b=0$ or $a-b=0$. If $a-b=0$ then $A=bJ_n$, where J_n is the $n \times n$ matrix in which each entry is one. Since J_n has rank one, the possibility $a=b$ contradicts hypothesis (v). Therefore $b=0$ and hence $A=aI_n$ as claimed.

*An invited paper. The preparation of this paper was supported in part by the U.S. Air Force Office of Scientific Research, under Grant 698–67.

**Present address: Department of Mathematics, University of California, Santa Barbara, Calif. 93106.

[1] Figures in brackets indicate the literature references at the end of this paper.

Before continuing, we explain a certain (known) device. Let $A' = DAD^{-1}$, where D is a diagonal matrix in which the (i, i) entry is $(-1)^i$ for $i = 1, 2, \ldots, n$. Applying a well-known formula [7] to A', we have

$$C_s(\text{adj } A') = \det A^{s-1} C^{n-s}(A'). \tag{1}$$

Here, as in [5], adj A' is the adjugate of A', $C_s(X)$ denotes the sth compound of X, and $C^{n-s}(X)$ denotes the corresponding supplementary compound. If $i_1, \ldots, i_s$ and $j_1, \ldots, j_s$ are two strictly increasing sequences of integers, the entry

$$(-1)^{i_1+\cdots+i_s+j_1+\cdots+j_s} \det A'(i_1, \ldots, i_s | j_1, \ldots, j_s)$$

of $C^{n-s}(A')$ equals

$$\det A[i_{s+1}, \ldots, i_n | j_{s+1}, \ldots, j_n].$$

Here $i_{s+1}, \ldots, i_n$ and $j_{s+1}, \ldots, j_n$ are the increasing sequences in $1, 2, \ldots, n$ complementary to $i_1, \ldots, i_s$ and $j_1, \ldots, j_s$, respectively. (Note: round brackets $A(|)$ indicate rows and columns deleted, and square brackets $A[|]$ indicate rows and columns retained.) It follows from these remarks and (1) that if A has all $(n-s) \times (n-s)$ principal minors equal, then adj A' has all $s \times s$ principal minors equal. It also follows that if all nonvanishing $(n-s) \times (n-s)$ nonprincipal minors of A have a common sign, then all nonvanishing nonprincipal $s \times s$ minors of adj A' have a common sign.

Second case: $r = n-2$. This is the largest value of r permitted and implies that A is nonsingular. Taking $s = 1$ and $s = 2$ in the discussion of the previous paragraph, we see that adj A' is scalar, and thus $(A')^{-1}$ is also scalar. But then A' is scalar, which implies that A is scalar. The proof is complete in case 2.

The general case: Let $1 < r < n-2$. We seek to prove that each $r \times r$ principal submatrix A_r of A is scalar. Since $r > 1$, it is quite easy to see from this fact that A is itself scalar.

If $\det A_r \neq 0$, we may pass a complete nested chain through A_r (see [5]), and so we secure $(r+1)$-square, $(r+2)$-square, and $(r+3)$-square principal submatrices A_{r+1}, A_{r+2}, A_{r+3} such that

$$A_r \subset A_{r+1} \subset A_{r+2} \subset A_{r+3} \tag{2}$$

with at least one of A_{r+2} or A_{r+3} invertible. If $\det A_r = 0$, then every r-square principal minor of A is singular, since these principal minors are all equal, and hence (as all $(r+1)$-square principal minors are equal) all $(r+1)$-square principal submatrices of A are nonsingular. Let A_{r+1} be an $(r+1)$-square principal submatrix containing A_r. Passing a complete nested chain through A_{r+1}, we obtain the nested chain (2), with at least one of A_{r+2} or A_{r+3} invertible.

If $\det A_{r+2} \neq 0$, we may apply case 2 to A_{r+2} and so conclude that A_{r+2} is scalar. Therefore A_r is also scalar.

If $\det A_{r+2} = 0$, then $\det A_{r+3} \neq 0$ and we let $B = \text{adj } A'_{r+3}$. Applying the identity (1) to A'_{r+3}, we see that B has all 2×2 principal minors equal, all nonvanishing 2×2 nonprincipal minors of common sign, all 3×3 principal minors equal, and all nonvanishing nonprincipal 3×3 minors of common sign. If we can show that B is scalar, it will follow that A_{r+3} is scalar, and hence that A_r is scalar. To prove that B is scalar, it will suffice to prove that each 2×2 principal submatrix B_2 of B is scalar. Since B is at least 5×5, we may embed B_2 in a nested chain

$$B_2 \subset B_3 \subset B_4 \subset B_5 \tag{3}$$

with at least one of B_4 or B_5 invertible. If $\det B_4 \neq 0$, an application of case 2 to B_4 shows that B_2 is scalar. Therefore we may suppose every 4×4 principal submatrix of B_5 containing B_2 is singular. If every 4×4 principal submatrix of B_5 is singular, then B_5 satisfies the hypotheses of Theorem 13 of [5] and hence

$$B_5 = aI_5 + bDJ_5D^{-1},$$

where $D = \operatorname{diag}(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5)$. Here $\epsilon_1, \ldots, \epsilon_5$ are each ± 1, and $a \neq 0$ as B_5 is nonsingular. We wish to show that $b = 0$. Since for $1 \leq i < j < k \leq 5$, we have $\det B_5[i, j|i, k] = \epsilon_i\epsilon_k ba$ and $\det B_5[i, k|j, k] = \epsilon_i\epsilon_j ab$, the sign condition on the nonprincipal 2×2 minors of B_5 yields (for $b \neq 0$) $\epsilon_1 = \epsilon_2 = \epsilon_3 = \epsilon_4 = \epsilon_5$. Thus, we have $B_5 = aI_5 + bJ_5$. But now $\det B_5[1, 2|1, 3] = ba$ and $\det B[2, 3|1, 2] = -ba$. The sign condition on the nonprincipal 2×2 minors of B_5 thus shows that $b = 0$. Hence B_5 is scalar, and therefore B_2 is scalar.

Remaining in our discussion of B_2 and B_5 is the case in which there is a $B_4 \natural B_2$ which is nonsingular. Suppose $B_2 = B_5[i, j|i, j]$ $(i < j)$. Then either B_4 overlaps B_2 in the (i, i) position or in the (j, j) position. Our B_4 satisfies the hypotheses of case 2. Hence B_4 is scalar, say $B_4 = aI_4$ with $a \neq 0$. Suppose B_4 and B_2 overlap in the (j, j) position. (The case in which B_4 and B_2 overlap in the (i, i) position can be obtained from this case by reversing the order of the rows and columns in B_5.) Let the (i, i) entry of B_5 be $a + b$. The equality of the principal 2×2 minors of B_5 shows that the nondiagonal entries in column i of B are $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_5$ with $x_1^2 = \ldots = x_{i-1}^2 = x_{i+1}^2 = \ldots = x_5^2 = ab$. If $i \geq 3$, $\det B_5[i, j|1, j] = ax_1 = -\det B_5[2, i|1, 2]$. If $i = 2$, $\det B_5[1, 5|1, 2] = ax_5 = -\det B_5[3, 5|2, 3]$. If $i = 1$, $\det B_5[2, 3|1, 2] = -ax_3 = -\det B_5[3, 4|1, 4]$. Therefore we must have one of $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_5$ zero, and hence $b = 0$. Thus all of $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_5, b$ are zero and hence B_5 is scalar and therefore B_2 is scalar.

The proof is now finished.

As immediate consequences of Theorem 1 we have Theorems 2 and 3.

THEOREM 2. *If a symmetric completely nonnegative matrix* A *of rank* $\geq r+2$ *has all* $r \times r$ *principal minors equal, and all* $(r+1) \times (r+1)$ *principal minors equal, then* A *is scalar.*

THEOREM 3. *A symmetric oscillation matrix cannot have all* $t \times t$ *principal minors equal for two consecutive values of* t.

It is not difficult to construct nontrivial examples of 3×3 completely positive matrices having all $t \times t$ principal minors equal, for one value of t. Therefore the hypothesis in Theorem 1 on the principal minors can be weakened only at the price of greatly strengthening the hypothesis concerning the nonprincipal minors.

THEOREM 4. *Let* A *be an* $n \times n$ *matrix with elements in a field* $\mathfrak{F}$. *Let* r *be a fixed integer,* $1 \leq r < n$. *Suppose:*

(i) *all* $r \times r$ *principal minors of* A *are equal;*

(ii) *all* $r \times r$ *nonprincipal minors of* A *are equal;*

(iii) *rank* $A \geq r+1$ *if* $r \neq 1$.

Then:

(a) *if* $r = 1$, $A = aI_n + bJ_n$, *where* $a, b \epsilon_i \mathfrak{F}$;

(b) *if* $r = n-1$, $A = D(aI_n + bJ_n)D^{-1}$, *where* $a, b \epsilon_i \mathfrak{F}$, *and* $D = \operatorname{diag}(-1, 1, -1, 1, \ldots, (-1)^i, \ldots, (-1)^n)$;

(c) *if* $1 < r < n-1$, $A = aI_n$ *is scalar.*

PROOF. First case: $r = 1$. This is trivial.

Second case: $r = n-1$. Here $\operatorname{adj} DAD^{-1}$ satisfies the conditions of case 1. Therefore $\operatorname{adj} DAD^{-1} = \alpha I_n + \beta J_n$. Thus $(DAD^{-1})^{-1} = \alpha' I_n + \beta' J_n$, and hence $DAD^{-1} = aI_n + bJ_n$. This yields the result in case 2.

General case: The assumptions (i) and (ii) of Theorem 4 are equivalent to

$$C_r(A) = aI_{\binom{n}{r}} + bJ_{\binom{n}{r}}, \text{ where } a, b \epsilon_i \mathfrak{F}. \tag{4}$$

Since $\operatorname{rank} C_r(A) = \binom{\operatorname{rank} A}{r} > 1$ (because $\operatorname{rank} A > r$), we see that $a \neq 0$. Let A_{r+1} be any principal $(r+1)$-square submatrix of A. Then $C_r(A_{r+1})$ is a principal submatrix of $C_r(A)$, so that (4) yields

$$C_r(A_{r+1}) = aI_{r+1} + bJ_{r+1}. \tag{5}$$

Let the eigenvalues of A_{r+1} be $\lambda_1, \ldots, \lambda_{r+1}$. These lie, of course, in an extension field of $\mathfrak{F}$. Then, as the eigenvalues of $aI_{r+1}+bJ_{r+1}$ are $a+(r+1)b$ (once), and a (r times), we may choose our notation for the eigenvalues of A_{r+1} so that

$$\lambda_1 \ldots \lambda_r = a+(r+1)b,\ \lambda_2\lambda_3 \ldots \lambda_{r+1} = a = \lambda_1\lambda_3 \ldots \lambda_{r+1}. \tag{6}$$

From (6) we see that all of $\lambda_1, \ldots, \lambda_{r+1}$ are nonzero and hence A_{r+1} is nonsingular.

Since A_{r+1} now satisfies the hypotheses of case 2, we see that A_{r+1} has the form

$$A_{r+1} = D(a_{r+1}I_{r+1}+b_{r+1}J_{r+1})D^{-1}, \tag{7}$$

where a_{r+1} and b_{r+1} are in $\mathfrak{F}$ and $D=\operatorname{diag}(-1, 1, -1, \ldots, (-1)^{r+1})$. From the formula (7) for A_{r+1}, it follows that the (i,j) element of A_{r+1} $(i<j<r+1)$ is the negative of the $(i,j+1)$ element of A_{r+1}. Applying this result to any A_{r+1} containing rows and columns $i, j, j+1$ of A, where $1 \leqslant i < j < n$, we see that the (i,j) element of A is the negative of the $(i,j+1)$ element of A, for $1 \leqslant i < j < n$. (An A_{r+1} exists containing rows and columns $i, j, j+1$ because $r+1 \geqslant 3$.) Next notice that

$$b = \det A[1, 2, \ldots, r|1, 2, \ldots, r-2, r+1, r+2], \tag{8}$$

since all nonprincipal r-square minors of A equal b. The last column of the minor in (8) is now known to be the negative of the second last column of this minor (8). Therefore $b=0$.

Hence each $C_r(A_{r+1}) = aI_{r+1}$. From the argument in case 2, it now follows that each A_{r+1} is scalar. Hence A is scalar. The proof is complete.

We remark that this proof can be shortened if $\mathfrak{F}$ is the real or complex number field.

We also remark that Theorem 4 is closely related to the question of the solvability of the matrix equation $C_r(X)=B$, where B is given. Recent results relating to this problem have been found by M. Marcus, M. Newman, A. Yaqub, H. Schwertdtfeger, W. Utz, and less recent results are to be found in papers by C. Ko, H. C. Lee, C. Yen, A. W. Wallace, D. E. Rutherford, A. C. Aitken, J. Williamson.

References

[1] Thompson, R. C., Principal submatrices of normal and Hermitian matrices, Illinois J. Math. **10**, 296–308 (1966).

[2] Thompson, R. C., and McEnteggert, P., Principal submatrices II: The upper and lower quadratic inequalities, J. Linear Algebra and Appl. **1**, 211–243 (1968).

[3] Thompson, R. C., Principal submatrices III: Linear inequalities, J. Res. NBS 72B (Math. Sci.) No. 1, 7–22 (1968).

[4] Thompson, R. C., Principal submatrices IV: On the independence of the eigenvalues of different principal submatrices, submitted to J. Linear Algebra and Appl.

[5] Thompson, R. C., Principal submatrices V: Some results concerning principal submatrices of arbitrary matrices, J. Res. NBS 72B (Math. Sci.) No. 2 (1968).

[6] Thompson, R. C., Principal submatrices VI: Cases of equality in certain linear inequalities submitted to J. Linear Algebra and Appl.

[7] Wedderburn, J. H. M., Lectures on matrices, Amer. Math. Soc. Colloq. Publ. **17**, 63–67 (New York, 1934).

(Paper 72B4–274)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 72B, No. 4, October–December 1968

On the Multipliers of the Dedekind Modular Function

Joseph Lehner*

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(September 20, 1968)

The Dedekind modular function is defined by $\eta(\tau)=e^{\pi i\tau/12}\prod_1^\infty(1-e^{2\pi i m\tau})$, $\operatorname{Im}\tau>0$, and satisfies the transformation equation $(c\tau+d)^{-1/2}\eta(A\tau)=v(A)\eta(\tau)$ for every $A\in\Gamma$, the modular group, where $v(A)$ is a complicated 24th root of unity depending on A. Let G be the set of all $A\in\Gamma$ for which $v(A)=1$. *Then G is not a group*, but there are groups that are subsets of G, e.g., $\{S^{24}\}$, where $S=\begin{pmatrix}1&1\\0&1\end{pmatrix}$. *Main Theorem.* Every subgroup of Γ that is a subset of G is cyclic. Moreover $G\subset\Gamma'$, the commutator subgroup of Γ.

Key Words: Cancellation; commutator subgroup; modular group; multiplier; word.

1. Let Γ be the matrix modular group, that is, $\Gamma=SL(2, Z)$, $Z=$ integers. The well-known Dedekind modular function $\eta(\tau)$ is defined by

$$\eta(\tau)=x^{1/24}\prod_1^\infty(1-x^m)=e^{\pi i\tau/12}\prod_1^\infty(1-e^{2\pi i m\tau}), \tag{1}$$

where, throughout, $x=e^{2\pi i\tau}$ and $\operatorname{Im}\tau>0$. The function η is a *modular form* of dimension $-1/2$ with *multiplier system* v; this means that

$$(c\tau+d)^{-1/2}\eta(A\tau)=v(A)\eta(\tau) \tag{2}$$

for each $A=\begin{pmatrix}a&b\\c&d\end{pmatrix}\in\Gamma$. Here $A\tau=(a\tau+b)/(c\tau+d)$ and $(c\tau+d)^{-1/2}$ is defined uniquely by setting, for each complex number u,

$$-\pi<\arg u\leqq\pi. \tag{3}$$

Moreover, the multiplier v is of modulus 1 and satisfies a consistency condition

$$v(AB)(c_{AB}\tau+d_{AB})^{1/2}=v(A)v(B)(c_AB\tau+d_A)^{1/2}(c_B\tau+d_B)^{1/2}, \tag{4}$$

where in general we write $A=\begin{pmatrix}a_A&b_A\\c_A&d_A\end{pmatrix}$. It is the multiplier system v that we wish to study.

The difficulty arises because η is a form of *fractional* dimension. For a form of dimension $-r$ the exponent r would occur in (4) instead of 1/2. Hence if r is integral, v is a homomorphism of Γ into the group of complex numbers of modulus 1. The subset G of Γ on which v is identically 1 is a subgroup of Γ.

This is no longer the case when r ceases to be an integer. *Then G is not a group*, but it does contain subgroups of Γ. For example, we see from (1) that $v(S^{24m})=1$, where

*University of Maryland and the National Bureau of Standards.

$$S=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}. \tag{5}$$

Hence the cyclic group $\{S^{24}\}$ is contained in G.

The main object of this paper is to establish the following conjecture of Rademacher:

THEOREM 1. *If H is a subgroup of Γ and a subset of G, then H is cyclic.*

We shall see that $G\subset\Gamma'$, the commutator subgroup of Γ. Thus H is a subgroup of Γ'. Since Γ' is free of rank 2, it follows from Theorem 1 that H is of infinite index in Γ' and is not normal. On the other hand, as well shall see later, G is of index 2 in Γ' in the sense that $\Gamma'=G\cup GS^{12}$ is a disjoint union and $v(AS^{12})=-1$ for every $A\epsilon G$.

2. We shall first treat a simpler case, namely, $\eta^2(\tau)$ with multipliers v^2. Here v^2 is a homomorphism and so, letting

$$G_1=\{A\epsilon\Gamma|v^2(A)=+1\},$$

we have

$$\Gamma'\subset G_1.$$

THEOREM 2. $G_1=\Gamma'$.

Since $[\Gamma:\Gamma']=12$—recall that we are dealing with the matrix groups—we have only to show that $[\Gamma:G_1]=12$. Since by (1) and (5), $v^2(S^h)=\exp \pi ih/6$, we see that $1, S, S^2, \ldots, S^{11}$ lie in different cosets of Γ/G_1. q.e.d.

We have already defined

$$G=\{A\epsilon\Gamma|v(A)=1\}.$$

Clearly $G\subset G_1$ and so

$$G\subset\Gamma'.$$

However, G is not a group. In fact, $-\begin{pmatrix}1 & 6\\ 0 & 1\end{pmatrix}\epsilon G$ but its inverse $-\begin{pmatrix}1 & -6\\ 0 & 1\end{pmatrix}$ does not. To discover what groups are contained in G, we must study (4) in detail.

Let us recall that we are writing modular matrices as

$$A=\begin{pmatrix}a_A & b_A\\ c_A & d_A\end{pmatrix},\ a_Ad_A-b_Ac_A=1.$$

If $c_A=0$, then $a_A=d_A=\pm 1$ and $A=\pm S^m$, $m=\pm b_A$. Define as usual

$$\operatorname{sgn} u=u/|u| \text{ for } u\neq 0.$$

For the systematic treatment of (4) we write $v(AB)=\sigma(A, B)v(A)v(B)$, so that

$$\sigma(A, B)=(c_AB\tau+d_A)^{1/2}(c_B\tau+d_B)^{1/2}(c_{AB}\tau+d_{AB})^{-1/2}. \tag{6}$$

Clearly $\sigma^2(A, B)=1$. Using the convention (3) together with (6) we obtain the following values:

$$v(I)=1,\ v(-I)=\exp(-\pi i/2),\ I=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

$$\sigma(S^m, A)=\sigma(A, S^m)=1 \tag{7}$$

$$\sigma(A, A^{-1})=1,\ c_A\neq 0;\ \sigma(A, A^{-1})=\operatorname{sgn} d_A,\ c_A=0.$$

The last equation shows that

$$v(A^{-1})=\bar{v}(A),\ c_A\neq 0;\ v(A^{-1})=\operatorname{sgn} d_A\cdot\bar{v}(A),\ c_A=0. \tag{8}$$

Next,

$$v(-A)=v(A)i\ \operatorname{sgn} c_A,\ c_A\neq 0;\ v(-A)=-v(A)i\ \operatorname{sgn} d_A,\ c_A=0. \tag{9}$$

Finally we have the following table, in which a missing entry means that the sign is irrelevant.

c_A	c_B	c_{AB}	$\sigma(A,B)$
+	+	+	+1
+	+	−	−1
+	−		+1
−	+		+1
−	−	−	+1
−	−	+	−1
0	+		sgn d_A
+	0		sgn d_B
0	−		+1
−	0		+1

(10)

Let

$$T=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix},\qquad R=\begin{pmatrix}0 & -1\\ 1 & 1\end{pmatrix}.$$

Note that

$$T^2=R^3=-I. \tag{11}$$

We introduce a symbol

$$V^\epsilon=-\epsilon R^\epsilon,\ \epsilon=\pm 1,\ V^1=V. \tag{12}$$

In V^ϵ, ϵ is not an exponent, and in fact we have

$$\begin{gathered}(V^\epsilon)^{-1}=-V^{-\epsilon},\ VV^{-1}=V^{-1}V=-I,\\ VV=-V^{-1},\ V^{-1}V^{-1}=V.\end{gathered} \tag{13}$$

Furthermore

$$TV=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}=S,\ TV^{-1}=\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}. \tag{14}$$

LEMMA 1. *Let*

$$A=TV^{\epsilon_1}\ \dots\ TV^{\epsilon_m}=\begin{pmatrix}a_A & b_A\\ c_A & d_A\end{pmatrix},\ \epsilon_1^2=1.$$

Then

$$a_A>0,\ b_A\geq 0,\ c_A\geq 0,\ d_A>0.$$

This is an immediate consequence of (14). The set of words of this form is therefore a semigroup.

We partition the elements of Γ, aside from $\pm T$ and $\pm V^\epsilon$, into 8 classes, as follows. Each element can be written uniquely as a power product of T and V, apart from sign. We set ($\epsilon_1^2=1$)

$$\begin{aligned}C_1&=\{A\epsilon\Gamma|A=TV^{\epsilon_1}\ \dots\ TV^{\epsilon_m}\},\\ C_2&=\{A\epsilon\Gamma|A=V^{\epsilon_1}T\ \dots\ V^{\epsilon_m}T\},\\ C_3&=\{A\epsilon\Gamma|A=TV^{\epsilon_1}\ \dots\ V^{\epsilon_m}T\},\\ C_4&=\{A\epsilon\Gamma|A=V^{\epsilon_1}T\ \dots\ TV^{\epsilon_m}\},\\ -C_i&=\{A\epsilon\Gamma|-A\epsilon C_i\},\ i=1,2,3,4.\end{aligned} \tag{15}$$

Since $(TV^\epsilon)^{-1}=V^{-\epsilon}T$, we observe that (in an obvious notation)

$$C_1^{-1}=C_2,\ C_3^{-1}=-C_3,\ C_3=C_1T,\ C_4^{-1}=-C_4,\ C_4=-TC_1 \tag{16}$$

Lemma 1 can now be restated: $A\epsilon C_1$ implies $a_A,\ d_A>0$; $b_A,\ c_A\geq 0$.

LEMMA 2. $v(S^m)=\exp(\pi im/12)$, $v(-S^m)=\exp(\pi i(m-6)/12)$, $v(\pm T)=\exp(\mp\pi i/4)$, $v(TV^\epsilon)=\exp(\pi i\epsilon/12)$.

The first values are immediate from the definition (1), and $v(T)$ is obtained from the classical transformation formula

$$\eta(T\tau)=(-i\tau)^{1/2}\eta(\tau).$$

Then use $R=TS$, $V^\epsilon=-\epsilon R^\epsilon$, and (8), (9), (10).

LEMMA 3. *When $A\epsilon C_1$, $v(A)=\exp(\pi i(\nu/12)$, where*

$$\nu=\nu(A)=\sum_{i=1}^{m}\epsilon_i.$$

The lemma is true when $m=1$, by Lemma 2. Suppose it is true for all words $\prod_1^h TV^{\epsilon_i}$, $h<m$, and let $A=\prod_1^m TV^{\epsilon_i}$. What we need to show is $v(ATV^\epsilon)=v(A)v(TV^\epsilon)$, in other words,

$$(*)\qquad \sigma(A,\ TV^\epsilon)=1.$$

Now by Lemma 1, $c_A\geq 0$. When $\epsilon=1$, $TV^\epsilon=S$ and when $c_A=0$, $A=S^m$ (by Lemma 1). In both cases (*) follows from (7). In the remaining case $c_A>0$ and $\epsilon=-1$, and (*) follows from line 1 of table (10) and (14), since

$$\begin{pmatrix} \cdot & \cdot \\ \geq 0 & >0 \end{pmatrix}\begin{pmatrix} 1 & \cdot \\ 1 & \cdot \end{pmatrix}=\begin{pmatrix} \cdot & \cdot \\ >0 & \cdot \end{pmatrix}.$$

3. We are now ready to begin the proof of Theorem 1. We assume there is a subgroup of Γ' that is contained in G and is not cyclic. This subgroup, then, contains a free group H of rank 2 as a subgroup and $H\subset G$. We denote by $\{A, B\}$ the group generated by A and B.

Every $A\epsilon\Gamma'$ falls into one and only one of the 8 classes of (15). If we write $A=\pm T^{\epsilon}V^{\epsilon_1}\cdots V^{\epsilon_m}T^{\epsilon'}$ with ϵ, $\epsilon'=0$, 1 and $\epsilon_i^2=1$, we can define the *length* of A to be m and denote it by $|A|$. From all $H=\{A, B\}$ select one, say $H_0=\{A_0, B_0\}$, such that $|A_0|+|B_0|$ is a minimum. This minimum is positive; call it ρ.

When we multiply two words, there may be internal cancellations. For example, suppose $A_0=\left(\prod_1^m TV^{\epsilon_i}\right)\cdot T$, $B_0=\prod_1^n TV^{\eta_i}$. Then if $\epsilon_m=-\eta_1$, the factor $V^{\epsilon_m}TTV^{\eta_1}$ will disappear from A_0B_0. Let $|A_0|\geq|B_0|$ and suppose B_0 cancels completely in A_0B_0. Then $|A_0B_0|=|A_0|-|B_0|$ and $|A_0B_0|+|B_0|=|A_0|<\rho$. Since $\{A_0, B_0\}=\{A_0B_0, B_0\}$, this is a contradiction. Hence there is a smallest interger $h\geq 0$ such that

$$w=\operatorname{sgn}(\epsilon_{m-h}+\eta_{1+h}) \tag{17}$$

is not zero. Similarly, by considering B_0A_0, $B_0^{-1}A_0$, and $B_0A_0^{-1}$, we deduce the existence of minimal nonnegative integers j, k, l such that

$$z=\operatorname{sgn}(\eta_{n-j}+\epsilon_{1+j}),\qquad x=\operatorname{sgn}(-\eta_{1+k}+\epsilon_{1+k}),\qquad y=\operatorname{sgn}(\eta_{n-l}-\epsilon_{m-l}) \tag{18}$$

are all nonzero. We shall use the quantities x, y, z, w systematically.

From now on we write A for A_0 and B for B_0.

Next we must develop rules that will enable us to multiply elements of the different classes. In the event of internal cancellations we shall have to take account of the relations (11) and (13). The following possibilities arise:

$$V^{\epsilon}TTV^{-\epsilon}=TV^{\epsilon}V^{-\epsilon}T=I,\ \epsilon=\pm 1 \tag{19}$$

$$V^{\epsilon}TTV^{\eta}=uV^{-u},\ TV^{\epsilon}V^{\eta}T=-uTV^{-u}T,\ \eta=\pm 1 \tag{20}$$

where $u=\operatorname{sgn}(\epsilon+\eta)\neq 0$. Suppose (19) occurs in the product AB. Then the exponent sum $\nu(AB)=\nu(A)+\nu(B)$. If (20) occurs, $\nu(AB)=\nu(A)+\nu(B)-\epsilon-\eta-u=\nu(A)+\nu(B)-3u$, since either $\epsilon=\eta=u=1$ or $\epsilon=\eta=u=-1$. Furthermore, we multiply the word AB by u in the first case of (20) and by $-u$ in the second case.

An example will make the technique clear. Consider BA, where $A\epsilon C_1$, $B\epsilon C_3$. Let $A=\prod_1^m TV^{\epsilon_i}$, $B=\left(\prod_1^n TV^{\eta_i}\right)\cdot T$. Here we have a cancellation of the type $V^{\eta}TTV^{\epsilon}$. Since z, defined in (18), is the sign of the first nonvanishing sum $\eta_n+\epsilon_1$, $\eta_{n-1}+\epsilon_2$, . . ., we have $BA=zDT$ with $D\epsilon C_1$ and $\nu(D)=\nu(B)+\nu(A)-3z$.

On the other hand suppose A is as before but $B\epsilon C_1$, $B=\prod_1^n TV^{\eta_i}$. Then in AB^{-1} there is a cancellation of the type $TV^{\epsilon}V^{-\eta}T$, so that $AB^{-1}=yDT$, since $-\chi=\operatorname{sgn}(\epsilon_{m-l}-\eta_{n-l})$. Also $\nu(D)=\nu(A)-\nu(B)+3y$.

One final rule. If $A\epsilon\pm C_3$ or $\pm C_4$, then for every appearance of A^{-1} in the word we are considering we must multiply by -1, since

$$\left\{\left(\pm\prod_1^m TV^{\epsilon_i}\right)\cdot T\right\}^{-1}=-\left(\pm\prod_1^m TV^{-\epsilon_{m-i}}\right)\cdot T \text{ and } \left\{\left(\pm\prod_1^{m-1} V^{\epsilon_i}T\right)\cdot V^{\epsilon_m}\right\}^{-1}=-\left(\pm\prod_1^m V^{-\epsilon_{m-i}}\right)\cdot V^{-\epsilon_1}.$$

4. We now proceed as follows. The generators A, B of H lie in the classes $\pm C_i$, $i=1, \ldots, 4$. For each possible pair A, B we shall exhibit a word $W(A, B)$ that does not belong to G.

Since $A, B \in G$, $\{A\}, \{B\} \subset G$.

LEMMA 4. *If $A \in -C_1$ or $A \in -C_2$, then $\{A\} \not\subset G$.*

Suppose first $A=-C_1$, $c_A=0$. Then $-A \in C_1$ and, by Lemma 1, $A=-S^m$. By Lemma 2, $v(A) = \exp(\pi i(m-6)/12)$, $v(A^2)=v(S^{2m}) = \exp(\pi i \cdot 2m/12)$. Hence $v(A)$, $v(A^2)$ cannot both equal 1, i.e., A and A^2 do not both belong to G. Next assume $c_A \neq 0$; it follows from Lemma 1 that $c_A < 0$. Hence $-A$ has all entries positive except $b_A \geq 0$ and so the c of $-A^2$ is positive. By line 1 of table 10, $\sigma(-A, -A)=1$. Therefore

$$v(A^2)=v(-A \cdot -A)=(v(-A))^2.$$

By (9), $v(A)$ and $v(-A)$ are not both real, hence $v(A)=v(A^2)=1$ is impossible.

Since $(-C_2)^{-1}=-C_1$, the result follows for $-C_2$.

LEMMA 5. *Let $D \in C_1$, $\epsilon=\pm 1$. Then*

$$\sigma(D, \epsilon T)=\sigma(\epsilon T, D)=\sigma(-T, DT)=1.$$

These results are proved by reference to table (10) and (7). By Lemma 1, $c_D \geq 0$, $d_D > 0$. Note that when $c_D > 0$, the c of $D \cdot \epsilon T$ and of ϵTD has sign ϵ.

The proof is now divided into cases: $A \epsilon \pm C_i$, $B \epsilon \pm C_j$, $i, j=1, \ldots, 4$, where we may obviously assume $j \geq i$. Lemma 4 shows we need not consider the classes $-C_1$ or $-C_2$. And in view of (16) we may disregard the classes $-C_3$ and $-C_4$ as well as C_2. There remain the cases: $A \epsilon C_1$, $B \epsilon C_1, C_3, C_4$; $A \epsilon C_3$, $B \epsilon C_3, C_4$; $A, B \epsilon C_4$. In each case we shall write the "cancellation rules" (cf. end of sec. 3) and exhibit a word in A and B for which $v=-1$.

$A \epsilon C_1$, $B \epsilon C_1$

	AB	AB^{-1}	$A^{-1}B$	$A^{-1}B^{-1}$	BA^{-1}	$B^{-1}A$
multiply by	1	y	$-x$	1	$-y$	x
add to exponent sum	0	$3y$	$3x$	0	$-3y$	$-3x$.

Here x, y are defined in (18) and

$$A=\prod_1^m TV^{\epsilon_i}, \qquad B=\prod_1^n TV^{\eta_i}.$$

Recall also (Lemma 6) that $\nu(A) \equiv \nu(B) \equiv 0 \pmod{24}$. The words for which $v=-1$ are as follows:

$y>0$; $B^{-1}ABA^{-1}BA^{-1}$

$y<0$, $x>0$; $B^{-1}AB^{-1}$

$y<0$, $x<0$; $B^{-1}ABA^{-1}B^{-1}A$.

$A \epsilon C_1$, $B \epsilon C_3$

	AB	AB^{-1}	$A^{-1}B$	$A^{-1}B^{-1}$	BA^{-1}	$B^{-1}A$	$B^{-1}A^{-1}$
multiply by	1	1	$-x$	$-z$	1	x	1
add to exponent sum	0	0	$3x$	$3z$	0	$-3x$	0.

Recall that when B^{-1} occurs a minus sign must be prefixed. Also $\nu(A) \equiv 0$, $\nu(B) \equiv 3 \pmod{24}$ by Lemma 6. The words are:

$$x > 0; \quad B^{-1}AB^{-1}$$

$$x < 0, \quad z > 0; \quad BA^{-1}B^{-1}AB$$

$$x < 0, \quad z < 0; \quad B^{-1}AB^{-1}A^{-1}B^{-1}.$$

$A \in C_1$, $B \in C_4$

	AB	$A^{-1}B$	$A^{-1}B^{-1}$	BA	BA^{-1}	$B^{-1}A$
multiply by	$-w$	1	1	1	$-y$	1
add to exponent sum	$-3w$	0	0	0	$-3y$	0.

Here $\nu(A) \equiv 0$, $\nu(B) \equiv -3 \pmod{24}$. When B^{-1} occurs, multiply by -1. The words are:

$$w > 0; \quad A^{-1}BAB$$

$$w < 0, \quad y > 0; \quad ABA^{-1}BA^{-1}$$

$$w < 0; \quad y < 0; \quad B^{-1}ABA^{-1}B^{-1}A$$

$A \in C_3$, $B \in C_3$

	AB	AB^{-1}	$A^{-1}B$	$A^{-1}B^{-1}$	BA	BA^{-1}	$B^{-1}A$	$B^{-1}A^{-1}$
multiply by	w	$-y$	$-x$	$-z$	z	y	x	$-w$
add to exponent sum	$-3w$	$3y$	$3x$	$3z$	$-3z$	$-3y$	$-3x$	$3w$.

When A^{-1} or B^{-1} occurs, prefix a minus sign; $\nu(A) \equiv \nu(B) \equiv 3 \pmod{24}$. The words are:

$$w < 0; \quad AB$$

$$z < 0; \quad BA$$

$$w, z > 0, \quad xy < 0; \quad ABA^{-1}BA^{-1}B$$

$$w, z > 0; \quad x, y > 0; \quad AB^{-1}A^{-1}B$$

$$w, z > 0; \quad x, y < 0; \quad A^{-1}B^{-1}A^{-1}BAB^{-1}.$$

$A \in C_3$, $B \in C_4$

In this case no internal cancellation ever occurs. We have, since $\nu(A) \equiv 3$, $\nu(B) \equiv -3 \pmod{24}$,

$$AB^{-1}A = -DT, \quad D \in C_1$$

with $\nu(D) \equiv 9 \pmod{24}$; the minus sign is caused by the presence of B^{-1}. Hence

$$v(AB^{-1}A) = v(D)v(-T) = \exp((\pi i/12)(9+3)) = -1.$$

$A \epsilon C_4$, $B \epsilon C_4$

	AB	AB^{-1}	$A^{-1}B$	BA	BA^{-1}	$B^{-1}A$	$B^{-1}A^{-1}$
multiply by	$-w$	y	x	z	$-y$	$-x$	w
add to exponent sum	$-3w$	$3y$	$3x$	$3z$	$-3y$	$-3x$	$3w$.

When A^{-1} or B^{-1} occurs, prefix a minus sign; $\nu(A) \equiv \nu(B) \equiv -3 \pmod{24}$. The words are:

$$x > 0, \quad y > 0; \quad A^{-1}B^{-1}A^{-1}BAB^{-1}$$

$$x > 0, \quad y < 0; \quad A^{-1}BA^{-1}$$

$$x < 0, \quad y > 0; \quad B^{-1}AB^{-1}$$

$$x < 0, \quad y < 0; \quad AB^{-1}A^{-1}BAB.$$

This completes the proof of Theorem 1.

5. We conclude with a list of generators for all cyclic subgroups of Γ' that are contained in G. Let $\{A\}$ be such a subgroup; then we may assume A is in one of the classes C_1, C_3, C_4 in view of Lemma 4 and (16).

Suppose $A \epsilon C_1$. Since $A \epsilon G$ we have $\nu(A) \equiv 0 \pmod{24}$ by Lemma 6. Suppose $c_A \neq 0$, i.e., $c_A > 0$. For $l \geq 1$, $A^l \epsilon C_1$ and $\nu(A^l) = l\nu(A)$. Thus the condition on $\nu(A)$ insures $v(A^l) = 1$, $l \geq 1$. Also the c of A^l is positive and so $v(A^{-l}) = \bar{v}(A^{-l}) = 1$, by (8). If $c_A = 0$ we have $A = S^m$, $24|m$ is necessary and sufficient for $\{S^m\}$ to lie in G. But $\nu(S^m) = m$. Thus $A \epsilon C_1$ generates a cyclic group contained in G if and only if

$$\nu(A) \equiv 0 \pmod{24}.$$

Next let $A \epsilon C_3$. By Lemma 6 we have $\nu(A) \equiv 3 \pmod{24}$. Also $c_A > 0$ always. Let $h \geq 0$ be the smallest integer such that

$$t = \operatorname{sgn}(\epsilon_{m-h} + \epsilon_{1+h}) \tag{21}$$

is not zero, where $A = \left(\prod_1^m TV^{\pi_i}\right) \cdot T$. If such an h does not exist, we would have $A^2 = -I$ and $\{A\}$ could not be a subgroup of the free group Γ'. Hence the existence of t is a necessary condition on A.

Multiplying A by A involves a concellation that has the effect of multiplying the word by t and adding $-3t$ to the exponent sum. Hence for $l \geq 1$,

$$A^l = t^{l-1}DT, \; D \epsilon C_1, \; \nu(D) \equiv 3l - 3(l-1)t \pmod{24}$$

and

$$v(A^l) = \exp\{(\pi i/12)(-3t^{l-1} + 3l - 3(l-1)t\}.$$

Hence $t=1$ is necessary and sufficient for $v(A^l)=1$, $l>0$. Since the c of $A^l=\pm DT$ is never zero, $v(A^{-l})=1$ also when $t=1$.

By similar reasoning we develop conditions in the case $A\epsilon C_4$ and then have the following

THEOREM 3. *Necessary and sufficient that a modular matrix A generate a group that is a subgroup of Γ' and a subset of G is that A satisfy the following conditions: A or A^{-1} lies in C_1, C_3, or C_4 and*

$$v(A')\equiv 0 \ (mod\ 24),\ if\ A.\epsilon C_1$$

$$v(A')\equiv 3 \ (mod\ 24)\ and\ t=1,\ if\ A'\epsilon C_3$$

$$v(A')\equiv -3 \ (mod\ 24)\ and\ t=-1,\ if\ A'\epsilon C_4.$$

Here A' is A or A^{-1} and t is defined by (21).

(Paper 72B4–275)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 72B, No. 4, October–December 1968

Minimax Error Selection of a Discrete Univariate Distribution With Prescribed Componentwise Bounds

A. J. Goldman and P. R. Meyers*

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(September 20, 1968)

The topic treated is that of finding a reproducible, plausible and computationally simple method of selecting a discrete frequency distribution with prescribed upper and lower bounds on its components. The problem is shown to be tractable when a minimax error selection criterion is employed, and "error" is measured by maximum absolute deviation between components. In this case one obtains a linear program of a special form admitting explicit solution. The vertices of the polyhedron of optimal solutions can also be found explicitly, and so their centroid can be calculated if unique specification is required.

Key Words: Linear programs; mathematical models; operations research.

1. Introduction

In the mathematical modelling efforts associated with an operations research study, one may well have only incomplete information on which to base a representation of the probabilities of the various outcomes of some pertinent chance event. Under these circumstances, one should of course examine the consequences of several alternative probability distributions, each consistent with the information at hand. It still seems desirable, however, to have a systematic and reproducible method for arriving at a *single* "nominal" distribution, to serve as a base-point for such sensitivity analyses.

This note works out the mathematics of one approach, based on a "minimax error" criterion, to the selection of a nominal distribution. The "incomplete information" is assumed to consist of upper and lower bounds on the individual terms of the probability distribution.

Let $\mathbf{L}$ and $\mathbf{U}$ be real n-vectors whose components satisfy

$$0 \leq L_i < U_i \leq 1. \tag{1.1}$$

A real n-vector $\mathbf{x}$ will be called a *probability vector* if it has nonnegative entries which sum to 1. We will be concerned with the set $P(\mathbf{L}, \mathbf{U})$ of probability vectors $\mathbf{x}$ whose components satisfy

$$L_i \leq x_i \leq U_i. \tag{1.2}$$

Our objective is to choose $\mathbf{x} \epsilon P(\mathbf{L}, \mathbf{U})$ to minimize

$$F(\mathbf{x}) = \max \{d(\mathbf{x}, \mathbf{y}) : \mathbf{y} \epsilon P(\mathbf{L}, \mathbf{U})\} \tag{1.3}$$

where d is the metric on n-space given by

$$d(\mathbf{x}, \mathbf{y}) = \max_i | x_i - y_i | . \tag{1.4}$$

*Present address: IBM World Trade Corporation, 821 United Nations Plaza, New York, New York 10017.

It would be interesting to consider the problem with the alternate metrics

$$d(\mathbf{x}, \mathbf{y}) = \Sigma_i (x_i - y_i)^2, \qquad d(\mathbf{x}, \mathbf{y}) = \Sigma_i \mid x_i - y_i \mid , \tag{1.5}$$

but this will not be done here. The common features of (1.4) and (1.5) are invariance under translation—used in the next section—and convexity in $\mathbf{y}$ for fixed $\mathbf{x}$. The latter property implies that the maximum in (1.3) occurs at an extreme point of $P(\mathbf{L}, \mathbf{U})$, and since the extreme points of this polyhedron can be found explicitly, $F(\mathbf{x})$ can be written as a discrete extremum. Thus the remainder of the problem—the minimization of F—has a strongly combinatorial flavor, but appears to be considerably harder for (1.5) than for (1.4).

The solution method is derived and justified in sections 2–4. In general there will be a convex polyhedron of optimal $\mathbf{x}$'s rather than a single one; section 5 proposes the centroid of this polyhedron's vertices as a plausible "representative" choice, and shows how these vertices can be calculated. If they are not too numerous, they might perhaps provide the basis for the sort of sensitivity analysis mentioned above. Section 6 shows how the solution method extends to the generalization

$$d(\mathbf{x}, \mathbf{y}) = \max_i w_i \mid x_i - y_i \mid$$

of (1.4), where the w_i are given positive numbers.

2. Simplification and Feasibility Analysis

From (1.3) and the form of d, it is apparent that the constraint set $P(\mathbf{L}, \mathbf{U})$ can be replaced by any of its translates without changing the problem. Our *first* simplification is a translation through $(-\mathbf{L})$, which replaces $P(\mathbf{L}, \mathbf{U})$ by the set

$$Q(S, \mathbf{V}) = \{\mathbf{x} : 0 \leq x_i \leq V_i \text{ (all } i), \Sigma_i x_i = S\} \tag{2.1}$$

where $\mathbf{V} = \mathbf{U} - \mathbf{L} > 0$ and $S = 1 - \Sigma_i L_i$. Now (1.3) is replaced by

$$F(\mathbf{x}) = \max \{d(\mathbf{x}, \mathbf{y}) : \mathbf{y} \epsilon Q(S, \mathbf{V})\}. \tag{2.2}$$

We turn next to a feasibility analysis, i.e., to determining when $Q(S, \mathbf{V})$ will be nonempty. The condition

$$0 \leq S \leq \Sigma_i V_i \tag{2.3}$$

is obviously necessary; we will show it is also sufficient. If equality holds throughout (2.3), then clearly $Q(S, \mathbf{V}) = \{\mathbf{0}\}$. If equality does not hold, then $(S/\Sigma_i V_i)\mathbf{V}$ is in $Q(S, \mathbf{V})$. From now on we assume (2.3) holds, in fact that

$$0 < S < \Sigma_i V_i, \tag{2.4}$$

since otherwise $Q(S, \mathbf{V})$ is a single point and the whole problem is trivial.

3. Evaluation of F

$F(\mathbf{x})$, as given by (1.4) and (2.2), is readily evaluated in terms of the quantities

$$M_i = \max \{y_i : \mathbf{y} \epsilon Q(S, \mathbf{V})\}, \tag{3.1}$$

$$m_i = \min \{y_i : \mathbf{y} \epsilon Q(S, \mathbf{V})\}. \tag{3.2}$$

Indeed, we have

$$F(\mathbf{x}) = \max_{\mathbf{y}} \max_i \max \{y_i - x_i, x_i - y_i\} = \max_i \max \{\max_{\mathbf{y}} (y_i - x_i), \max_{\mathbf{y}} (x_i - y_i)\}$$

or finally

$$F(\mathbf{x}) = \max_i \max \{M_i - x_i, x_i - m_i\}. \tag{3.3}$$

We next find explicit expressions for M_i and m_i, so that (3.3) is also explicit. It will be shown that

$$M_i = \min \{V_i, S\}, \tag{3.4}$$

$$m_i = \max \{0, S - \Sigma_{j \neq i} V_j\}. \tag{3.5}$$

If M_i^* and m_i^* denote the right-hand sides of (3.4) and (3.5), then clearly any $\mathbf{y} \epsilon Q(S, \mathbf{V})$ satisfies $m_i^* \leq y_i \leq M_i^*$. It suffices to show that $y_i = M_i^*$ for some $\mathbf{y} \epsilon Q(S, \mathbf{V})$, and that $y_i = m_i^*$ for some $\mathbf{y} \epsilon Q(S, \mathbf{V})$.

For the first purpose, if $V_i \leq S$ we note from (2.4) that

$$0 \leq \theta = (S - V_i)/\Sigma_{j \neq i} V_j < 1,$$

and set

$$y_i = V_i, \; y_j = \theta V_j \text{ for } j \neq i.$$

If $S < V_i$, we set

$$y_i = S, \; y_j = 0 \qquad \text{for } j \neq i.$$

For the second purpose, if $S \geq \Sigma_{j \neq i} V_j$ we set

$$y_i = S - \Sigma_{j \neq i} V_j, \; y_j = V_j \text{ for } j \neq i.$$

If $S < \Sigma_{j \neq i} V_j$ we set

$$y_i = 0, \; y_j = V_j S / \Sigma_{k \neq i} V_k \text{ for } j \neq i.$$

This completes the proof of (3.4) and (3.5).

We now make a second simplification, namely translation of the constraint set $Q(S, \mathbf{V})$, through $(-\mathbf{m})$ where $\mathbf{m}$ is the vector with components m_i. The result is a new constraint set $Q(S', \mathbf{V}')$, where

$$S' = S - \Sigma_i m_i, \; \mathbf{V}' = \mathbf{V} - \mathbf{m}.$$

However, the primes will be dropped. We now have all $m_i = 0$, so that

$$F(\mathbf{x}) = \max_i \{M_i - x_i, x_i\}, \tag{3.6}$$

$$S \geq \Sigma_{j \neq i} V_j \qquad (\text{all } i). \tag{3.7}$$

Note that (3.4) still holds.

4. Minimization of F

The minimization of F over $Q(S, \mathbf{V})$ can be viewed as a linear program, namely to select z and $\mathbf{x}$ so as to minimize z, subject to $\mathbf{x} \epsilon Q(S, \mathbf{V})$ and to

$$z \geq M_i - x_i; \; z \geq x_i \qquad (\text{all } i).$$

321-453 OL-69—2

These last constraints are equivalent to

$$M_i - z \leq x_i \leq z,$$

so that the linear program's full set of constraints reads

$$\max \{0, M_i - z\} \leq x_i \leq \min \{z, V_i\} \qquad (\text{all } i), \tag{4.1}$$

$$\Sigma_i x_i = S. \tag{4.2}$$

For fixed z, the existence of an $\mathbf{x}$ satisfying (4.1) and (4.2) clearly implies that

$$\Sigma_i \max \{0, M_i - z\} \leq S \leq \Sigma_i \min \{z, V_i\}, \tag{4.3}$$

$$z \geq \tfrac{1}{2} \max_i M_i > 0. \tag{4.4}$$

We now show that the converse also holds. Note first that (3.4) and (4.4) yield

$$V_i \geq M_i > M_i - z,$$

which together with (4.4) yields

$$\max \{0, M_i - z\} \leq \min \{z, V_i\} \qquad (\text{all } i). \tag{4.5}$$

If equality holds in the second part of (4.3), then (4.1) and (4.2) are satisfied by taking

$$x_i = \min \{z, V_i\} \qquad (\text{all } i).$$

If strict inequality holds, then $k = n$ has the property

$$\Sigma_{j \leq k} \min \{z, V_j\} + \Sigma_{j > k} \max \{0, M_j - z\} > S, \tag{4.6}$$

but $k = 0$ does not, and so there is a smallest $k \in \{1, 2, \ldots, n\}$ with this property; for that k,

$$\Sigma_{j < k} \min \{z, V_j\} + \Sigma_{j \geq k} \max \{0, M_j - z\} \leq S. \tag{4.7}$$

We now set

$$x_j = \min \{z, V_j\} \qquad (\text{all } j < k),$$

$$x_j = \max \{0, M_j - z\} \qquad (\text{all } j > k),$$

$$x_k = S - \Sigma_{j \neq k} x_j;$$

that (4.1) is satisfied for $i = k$ follows from (4.6) and (4.7).

The problem has now been reduced to minimizing z subject to (4.3) and (4.4). For this purpose, renumber so that

$$V_1 \geq V_2 \geq \ldots \geq V_n > V_{n+1} = 0, \tag{4.8}$$

which by (3.4) implies

$$M_1 \geq M_2 \geq \ldots \geq M_n > M_{n+1} = 0. \tag{4.9}$$

If z^* is the minimum value of $z>0$ satisfying the first half of (4.3),

$$\Sigma_i \max\{0, M_i - z\} \leq S, \tag{4.10}$$

and z^{**} is the minimum value of $z>0$ satisfying the second half,

$$\Sigma_i \min\{z, V_i\} \geq S, \tag{4.11}$$

then the desired minimum value of z is

$$z_{\min} = \max\{z^*, z^{**}, \tfrac{1}{2}\min\{S, V_1\}\}. \tag{4.12}$$

Thus it suffices to determine z^* and z^{**}.

The determination of z^* goes as follows. Let

$$M_j^* = \Sigma_1^j M_i - jM_j. \tag{4.13}$$

Then $M_1^* = 0 < S$, and by (2.4) we have $M_{n+1}^* \geq S$. Moreover, the calculation

$$M_{j+1}^* - M_j^* = j(M_j - M_{j+1}) \geq 0$$

shows that the sequence $\{M_j^*\}_1^{n+1}$ is nondecreasing. Thus there is a unique $J \epsilon \{1, 2, \ldots, n\}$ such that

$$0 = M_1^* \leq M_2^* \leq \ldots \leq M_J^* < S \leq M_{J+1}^* \leq \ldots \leq M_{n+1}^*. \tag{4.14}$$

Now if $0 < z < M_{J+1}$, then

$$\Sigma_i \max\{0, M_i - z\} \geq \Sigma_1^{J+1} M_i - (J+1)z > M_{J+1}^* \geq S,$$

so that z does not satisfy (4.10). And if $M_J \geq z \geq M_{J+1}$, then (4.10) becomes

$$\Sigma_1^J M_i - Jz \leq S, \tag{4.15}$$

which is equivalent to

$$z \geq z^* = (\Sigma_1^J M_i - S)/J. \tag{4.16}$$

By use of (4.14), the value of z^* proposed in (4.16) is easily verified to satisfy $M_J \geq z^* \geq M_{J+1}$, and so is indeed the smallest $z>0$ satisfying (4.10).

The determination of z^{**} goes as follows. Let

$$V_k^{**} = kV_k + \Sigma_{k+1}^n V_i. \tag{4.17}$$

Then by (2.4), we have $V_1^{**} > S$, and $0 = V_{n+1}^{**} < S$. Also, the calculation

$$V_{k+1}^{**} - V_k^{**} = k(V_{k+1} - V_k) \leq 0$$

shows that the sequence $\{V_k^{**}\}_1^{n+1}$ is nonincreasing. Thus there is a unique $K \epsilon \{1, 2, \ldots, n\}$ such that

$$V_1^{**} \geq V_2^{**} \geq \ldots \geq V_K^{**} \geq S > V_{K+1}^{**} \geq \ldots \geq V_{n+1}^{**}. \tag{4.18}$$

Now if $z \leq V_{K+1}$, then

$$\Sigma_i \min\{z, V_i\} \leq (K+1)z + \Sigma_{K+2}^n V_i \leq V_{K+1}^{**} < S,$$

so that z does not satisfy (4.11). And if $V_K \geq z > V_{K+1}$, then (4.11) becomes

$$Kz + \Sigma_{K+1}^{n} V_i \geq S,$$

which is equivalent to

$$z \geq z^{**} = (S - \Sigma_{K+1}^{n} V_i)/K. \tag{4.19}$$

By use of (4.18), the value of z^{**} proposed in (4.19) is easily verified to satisfy $V_K \geq z^{**} > V_{K+1}$, and so is indeed the smallest $z > 0$ satisfying (4.11).

The process for finding K, and thus z^{**}, can be shortened as follows. First observe that the constraint set $Q(S, \mathbf{V})$ can clearly be replaced by $Q(S, \mathbf{M})$, where the components M_i of $\mathbf{M}$ are given by (3.4). Thus V_k's are replaced by M_k's, and V_k^{**}'s by M_k^{**}'s. From (4.13) and (4.17) we have

$$M_j^* + M_j^{**} = \Sigma_1^n M_i,$$

so that (4.18) is equivalent to

$$M_K^* \leq \Sigma_1^n M_i - S < M_{K+1}^*,$$

and K can be evaluated using the same sequence $\{M_j^*\}_1^n$ used to locate J. This sequence can be generated using the initial condition $M_1^* = 0$ and the recursion

$$M_{j+1}^* = M_j^* + j(M_j - M_{j+1}).$$

To conclude this section, we recapitulate the solution process developed above. The input data are the n-vectors $\mathbf{L}$ and $\mathbf{U}$, satisfying $0 \leq \mathbf{L} < \mathbf{U}$. (If some $L_i = U_i$, we set $x_i = L_i$, drop the ith component from consideration, and rescale $\mathbf{x}$ by $(1 - L_i)^{-1}$.)

STEP 1: Set $V_i = U_i - L_i$ for $1 \leq i \leq n$. Set $S = 1 - \Sigma_i L_i$.

STEP 2: If $S < 0$ or $S > \Sigma_i V_i$, STOP; the problem is infeasible. If $S = 0$ set $\mathbf{x} = \mathbf{0}$ and go to Step 10; if $S = \Sigma_i V_i$ set $\mathbf{x} = \mathbf{V}$ and go to Step 10.

STEP 3: Form the vector $\mathbf{m}$ with conponents $m_i = \max\{0, S - \Sigma_{j \neq i} V_j\}$. Replace $\mathbf{V}$ by $\mathbf{V} - \mathbf{m}$, and S by $S - \Sigma_i m_i$.

STEP 4: Renumber the components of $\mathbf{V}$ to be in nonincreasing order. Record the corresponding permutation π of $\{1, 2, \ldots, n\}$. Set $V_{n+1} = 0$.

STEP 5: Calculate $M_i = \min\{V_i, S\}$ for $1 \leq i \leq n+1$.

STEP 6: Calculate M_j^*, from the initial condition $M_1^* = 0$ and the recursion $M_{j+1}^* = M_j^* + j(M_j - M_{j+1})$, until a first one M_{J+1}^* is reached which is $\geq S$ and a first one M_{K+1}^* is reached which is $> \Sigma_1^n M_i - S$. Set $z^* = (\Sigma_1^J M_i - S)/J$ and $z^{**} = (S - \Sigma_{K+1}^n M_i)/K$.

STEP 7: Form $z = \max\{z^*, z^{**}; \frac{1}{2}M_1\}$. (This corresponds to the $z_{\min}$ of the previous text.)

STEP 8: (Assumes any optimal $\mathbf{x}$ will do.) Calculate S_j^*, from the initial condition $S_n^* = \Sigma_1^n \min\{z, V_i\}$ and the backwards recursion

$$S_{j-1}^* = S_j^* + \max\{0, M_j - z\} - \min\{z, V_j\},$$

until a first one S_{k-1}^* is reached which is $\leq S$. Define $\mathbf{x}$ by

$$\begin{aligned} x_j &= \min\{z, V_j\} && (\text{all } j < k), \\ x_j &= \max\{0, M_j - z\} && (\text{all } j > k), \\ x_k &= S - \Sigma_{j \neq k} x_j. \end{aligned}$$

STEP 9: Permute the components of $\mathbf{x}$ by π^{-1}. Then replace $\mathbf{x}$ by $\mathbf{x} + \mathbf{m}$.

STEP 10: Replace $\mathbf{x}$ by $\mathbf{x} + \mathbf{L}$.

5. The Centroid Solution

Once $z_{\min}$ has been found, the optimal $\mathbf{x}$'s are defined by the constraints (4.1) and (4.2), which we repeat here,

$$\Sigma_i x_i = S, \tag{5.1}$$

$$a_i \leq x_i \leq b_i \qquad \text{(all } i\text{)} \tag{5.2}$$

where $a_i = \max\{0, M_i - z_{\min}\}$ and $b_i = \min\{z_{\min}, V_i\}$. One member of this convex polyhedron of solutions can be found by the construction used to show that (4.3) and (4.4) implied the existence of a solution to (4.1) and (4.2). But this is somewhat arbitrary. In view of the original intention to fasten on a single $\mathbf{x}$, it seems less arbitrary (at least to the writers) to select the *centroid* of the vertices of the polyhedron. Fortunately, these extreme points can be found explicitly. In what follows we assume

$$\Sigma_i a_i < S < \Sigma_i b_i,$$

thus excluding the trivial cases in which the polyhedron degenerates to a single point.

The polyhedron lies in the $(n-1)$-dimensional hyperplane (5.1); its extreme points are characterized as intersections of (5.1) with some $n-1$ of the $2n$ "bounding hyperplanes" $x_i = a_i$ and $x_i = b_i$. Such a point of intersection, $\mathbf{x}(j, A)$, is associated to each $j \in \{1, \ldots, n\}$ and each of the 2^{n-1} subsets A of $\{1, 2, \ldots, n\} - \{j\}$. The coordinates of $\mathbf{x}(j, A)$ are given by

$$x_i = a_i \qquad i \in A,$$

$$x_i = b_i \qquad i \in \{1, 2, \ldots, n\} - \{j\} - A,$$

$$x_j = S - \Sigma_{i \neq j} x_i.$$

The extreme points of the polyhedron are precisely those $\mathbf{x}(j, A)$ for which

$$a_j \leq x_j \leq b_j.$$

6. The Weighted Version

We now generalize the objective function (1.3) to

$$F(\mathbf{x}) = \max\{\max_i w_i |x_i - y_i| : \mathbf{y} \in P(\mathbf{L}, \mathbf{U})\} \tag{6.1}$$

where $\{w_i\}_1^n$ is a set of positive "weights" with $\Sigma_i w_i = 1$. This corresponds to the case in which the accuracy, with which $\mathbf{x}$ approximates the "true" distribution, is more important for some components than for others.

The reasoning of sections 2 and 3 go through unchanged, with occasional insertions of "w_i" in appropriate formulas. We find that (4.1) is replaced by

$$\max\{0, M_i - z/w_i\} \leq x_i \leq \min\{z/w_i, V_i\} \qquad \text{(all } i\text{)} \tag{6.2}$$

so that (4.3) and (4.4) are replaced by

$$\Sigma_i \max\{0, M_i - z/w_i\} \leq S \leq \Sigma_i \min\{z/w_i, V_i\}, \tag{6.3}$$

$$z \geq \tfrac{1}{2} \max_i M_i w_i. \tag{6.4}$$

The problem is now that of minimizing z subject to (6.3) and (6.4). As before, we write

$$z_{\min} = \max \{z^*, z^{**}, \tfrac{1}{2} \max_i M_i w_i\}, \tag{6.5}$$

where now z^* is the minimum value of $z \geq 0$ satisfying

$$\Sigma_i \max \{0, M_i - z/w_i\} \leq S \tag{6.6}$$

and z^{**} is the minimum value of $z \geq 0$ satisfying

$$\Sigma_i \min \{z/w_i, V_i\} \geq S. \tag{6.7}$$

The renumbering now is to be such that

$$V_1 w_1 \geq V_2 w_2 \geq \ldots \geq V_n w_n > 0 \tag{6.8}$$

which by (3.1) implies a similar condition for $\{M_i w_i\}_1^n$. As before we set $V_{n+1} = 0$ (implying $M_{n+1} = 0$), and now also choose any $w_{n+1} > 0$.

The determination of z^* goes as follows. Let

$$M_j^* = \Sigma_1^j M_i - w_j M_j \Sigma_1^j (1/w_i).$$

Then $M_1^* = 0 < S$, and by (2.4) we have $M_{n+1}^* \geq S$. The calculation

$$M_{j+1}^* - M_j^* = (w_j M_j - w_{j+1} M_{j+1}) \Sigma_1^j (1/w_i) \geq 0$$

shows that sequence $\{M_j^*\}_1^{n+1}$ is nondecreasing. Thus there is a unique $J \epsilon \{1, 2, \ldots, n\}$ such that

$$0 = M_1^* \leq M_2^* \leq \ldots \leq M_J^* < S \leq M_{J+1}^* \leq \ldots \leq M_{n+1}^*. \tag{6.9}$$

Now if $0 < z < M_{J+1} w_{J+1}$, then

$$\Sigma_i \max \{0, M_i - z/w_i\} \geq \Sigma_1^{J+1} (M_i - z/w_i) > M_{J+1}^* \geq S,$$

so that z does not satisfy (6.6). And if $M_J w_J \geq z \geq M_{J+1} w_{J+1}$, then (6.6) becomes

$$\Sigma_1^J (M_i - z/w_i) \leq S,$$

which is equivalent to

$$z \geq z^* = (\Sigma_1^J M_i - S)/\Sigma_1^J (1/w_i). \tag{6.10}$$

By use of (6.9), the value of z^* proposed in (6.10) is easily verified to satisfy $M_J w_J \geq z^* \geq M_{J+1} w_{J+1}$, and so is indeed the smallest $z > 0$ satisfying (6.6).

The determination of z^{**} is similar, involving

$$V_k^{**} = w_k V_k \Sigma_1^k (1/w_i) + \Sigma_{k+1}^n V_i,$$

$$V_1^{**} \geq V_2^{**} \geq \ldots \geq V_K^{**} \geq S > V_{K+1}^{**} \geq \ldots \geq V_{n+1}^*,$$

$$z^{**} = (S - \Sigma_{K+1}^{n} V_i)/\Sigma_1^K (1/w_i). \tag{6.11}$$

(Paper 72A4–276)

JOURNAL OF RESEARCH of the National Bureau of Standards – B. Mathematical Sciences
Vol. 72B, No. 4, October–December 1968

Minimax Error Selection of a Discrete Univariate Distribution With Prescribed Componentwise Ranking

A. J. Goldman

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(September 27, 1968)

The topic treated is that of finding a reproducible, plausible and computationally simple method of selecting a discrete frequency distribution with a prescribed ranking of its components. The problem is shown to be tractable when a minimax error selection criterion is employed, and "error" is measured by maximum absolute deviation between components. The vertices of the polyhedron of optimal solutions can also be found explicitly, and so their centroid can be calculated if unique specification is required.

Key Words: Mathematical models; minimax estimation; probability distribution.

1. Introduction

In the mathematical modeling efforts associated with an operations research study, one may well have only incomplete information on which to base a representation of the probabilities of the various outcomes of some pertinent chance event. Under these circumstances, one should of course examine the consequences of several alternative probability distributions, each consistent with the information at hand. It still seems desirable, however, to have a systematic and reproducible method for arriving at a *single* "nominal" distribution, to serve as a base-point for such sensitivity analyses.

This note works out the mathematics of one approach, based on a "minimax error" criterion, to the selection of a nominal distribution. The "incomplete information" is assumed to consist of a *ranking* of the individual terms of the probability distribution.[1] A previous paper[2] considered the case in which the given information consists of prescribed upper and lower bounds on these terms.

Let P be the polyhedron of real n-vectors $\mathbf{x}$ (where $n > 1$) whose components x_i satisfy

$$0 \leq x_1 \leq x_2 \leq \ldots \leq x_n, \tag{1.1}$$

$$\Sigma_i x_i = 1. \tag{1.2}$$

Our objective is to choose $\mathbf{x} \epsilon P$ to minimize

$$F(\mathbf{x}) = \max \{d(\mathbf{x}, \mathbf{y}) : \mathbf{y} \epsilon P\}, \tag{1.3}$$

where d is the metric on n-space given by

$$d(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i|. \tag{1.4}$$

[1] Study of this situation was suggested by J. McLynn of Davidson, Talbird and McLynn, Inc.

[2] A. J. Goldman and P. R. Meyers, Minimax error selection of a discrete univariate distribution with prescribed componentwise bounds, J. Res. NBS **72B** (Math. Sci.) No. 4 263-271 (1968).

Other metrics may be investigated subsequently.

The following section accomplishes the explicit evaluation of $F(\mathbf{x})$ in highly tractable form. The minimization of $F(\mathbf{x})$ over P, which begins section 3, is then trivial. It turns out, however, that for $n > 2$ there will be a convex polyhedron of optimal $\mathbf{x}$'s rather than a single one; section 3 goes on to propose the centroid of this polyhedron's vertices as a plausible "representative" choice, and includes a determination of both vertices and centroid.

2. Evaluation of F

The particular metric (1.4) has the pleasant property that $F(\mathbf{x})$ – for *any* compact set P, not only the one defined by (1.1) and (1.2) – can readily be evaluated in terms of the quantities

$$M_i = \max \{y_i : \mathbf{y} \epsilon P\}, \tag{2.1}$$

$$m_i = \min \{y_i : \mathbf{y} \epsilon P\}. \tag{2.2}$$

Indeed, we have

$$F(x) = \max_y \max_i \max (y_i - x_i, x_i - y_i)$$

$$= \max_i \max \{\max_y (y_i - x_i), \max_y (x_i - y_i)\},$$

or finally

$$F(x) = \max_i \max \{M_i - x_i, x_i - m_i\}. \tag{2.3}$$

For the particular polyhedron P which figures here, it is easily verified that

$$M_i = 1/(n+1-i) \qquad (1 \leqslant i \leqslant n), \tag{2.4}$$

$$m_i = 0 \qquad (1 \leqslant i < n), \tag{2.5}$$

$$m_n = 1/n. \tag{2.6}$$

Thus we have

$$F(\mathbf{x}) = \max \{\max_i \{1/(n+1-i) - x_i\}, x_{n-1}, x_n - 1/n\}.$$

The term corresponding to $i = n$ is

$$1 - x_n = \textstyle\sum_1^{n-1} x_i \geqslant x_{n-1},$$

so we can replace the last equation by

$$F(\mathbf{x}) = \max \{\max_i \{1/(n+1-i) - x_i\}, x_n - 1/n\}. \tag{2.7}$$

Next, suppose the term with $i = n-1$ yields the maximum in (2.7). Then we should have

$$1/2 - x_{n-1} \geqslant x_n - 1/n, \tag{2.8}$$

$$1/2 - x_{n-1} \geqslant 1 - x_n. \tag{2.9}$$

Addition yields

$$1-2x_{n-1} \geq 1-1/n,$$

which is equivalent to $x_{n-1} \leq 1/2n$. Thus $x_i \leq 1/2n$ for $1 \leq i < n-1$, and so

$$1-x_{n-1}-x_n = \sum_1^{n-2} x_i \leq (n-2)/2n.$$

This and (2.8) imply

$$(n+2)/2n - x_{n-1} \leq x_n = 1/n + (x_n - 1/n) \leq 1/n + (1/2 - x_{n-1}) = (n+2)/2n - x_{n-1}. \tag{2.10}$$

Since equality holds in (2.10), it must hold throughout the preceding sequence of equalities. In particular it holds in (2.8) and (2.9), so that the maximum in (2.7) does not occur *uniquely* for the term with $i = n-1$, and hence this term can be deleted from (2.7):

$$F(\mathbf{x}) = \max\{\max_{i<n-1}\{1/(n+1-i) - x_i\},\ 1-x_n,\ x_n - 1/n\}. \tag{2.11}$$

For $n=2$, (2.11) gives

$$F(\mathbf{x}) = \max\{1-x_2,\ x_2 - 1/2\}. \tag{2.12}$$

We next assume $n \geq 3$.

Now suppose the maximum in (2.11) holds for the term corresponding to some $i < n-1$. Then we should have

$$1/(n+1-i) - x_i \geq 1 - x_n, \tag{2.13}$$

$$1/(n+1-i) - x_i \geq x_n - 1/n. \tag{2.14}$$

Addition yields

$$2/(n+1-i) - 2x_i \geq 1 - 1/n,$$

or equivalently

$$\begin{aligned} 2x_i &\leq 2/(n+1-i) + 1/n - 1 \\ &= [2n + (n+1-i) - n(n+1+i)]/n(n+1-i) \\ &= [1 + 2n - n^2 + (n-1)i]/n(n+1-i) \\ &= [2 - (n-1)^2 + (n-1)i]/n(n+1-i) \\ &= [2 - (n-1)(n-1-i)]/n(n+1-i). \end{aligned}$$

Since $x_i \geq 0$, we must have

$$2 \geq (n-1)(n-1-i). \tag{2.15}$$

But $n-1 \geq 2$ and $n-i > 1$. Thus this situation is impossible for $n > 3$, i.e.,

$$F(\mathbf{x}) = \max\{1-x_n,\ x_n - 1/n\} \qquad (n > 3). \tag{2.16}$$

And it is possible for $n=3$ only when $i=n-2=1$, in which case equality holds in (2.15), hence throughout the preceding inequalities, so that the maximum in (2.11) does not hold *uniquely* for the term corresponding to $i=1$. We have now shown that, for *all* $n>1$,

$$F(\mathbf{x})=\max\{1-x_n,\ x_n-1/n\}. \tag{2.17}$$

3. Determination of Optimal Distributions

The minimization of $F(\mathbf{x})$ over P is trivial in view of (2.17), which can be rewritten

$$\begin{aligned} F(\mathbf{x})&=1-x_n \text{ for } 0\leqq x_n\leqq (n+1)/2n,\\ F(\mathbf{x})&=x_n-1/n \text{ for } (n+1)/2n\leqq x_n\leqq 1. \end{aligned} \tag{3.1}$$

We see at once that

$$F_{\min}=(n-1)/2n, \tag{3.2}$$

and that the minimizing $\mathbf{x}\epsilon P$ are precisely those with

$$x_n=(n+1)/2n. \tag{3.3}$$

One such $\mathbf{x}$ is given by

$$x_i=1/2n \qquad \text{for } 1\leqq i\leqq n-1. \tag{3.4}$$

But this choice is somewhat arbitrary, and moreover honors the original ranking (for $x_1, \ldots, x_{n-1}$) only in a technical sense. In view of the original intention to fasten on a single $\mathbf{x}$, it seems (to the writer) less arbitrary to select the *centroid* of the vertices of the polyhedron of F — minimizing $\mathbf{x}\epsilon P$. This centroid is obtained by adjoining an nth component $(n+1)/2n$, as in (3.3), to the centroid of the polyhedron Q in $(n-1)$-space consisting of those $\mathbf{z}$ with

$$0\leqq z_1\leqq \ \ldots\ \leqq z_{n-1}\leqq (n+1)/2n, \tag{3.5}$$

$$\Sigma_1^{n-1}z_i=1-(n+1)/2n=(n-1)/2n. \tag{3.6}$$

Note that in view of (3.6), the last inequality in (3.5) can be dropped.

To determine the vertices of Q, note that a nonsingular linear transformation preserves extreme points of polyhedra. This applies in particular to the transformation associated with the substitutions

$$\begin{aligned} u_1&=z_1,\\ u_i&=z_i-z_{i-1} \qquad (2\leqq i\leqq n-1), \end{aligned}$$

which convert Q into the polyhedron Q' defined in $\mathbf{u}$-space by

$$u_i\geqslant 0 \qquad (1\leqq i\leqq n-1), \tag{3.7}$$

$$\Sigma_{i=1}^{n-1}(n-i)u_i=(n-1)/2n. \tag{3.8}$$

Q' consists of the intersection of the hyperplane H, defined by (3.8), with the nonnegative orthant.

Thus its extreme points are the intersections of H with the $(n-1)$ positive coordinate axes, i.e., the $(n-1)$ points $\mathbf{u}^{(j)}$ given by

$$u_i^{(j)} = \delta_{ij}(n-1)/2n(n-j). \tag{3.9}$$

The corresponding extreme points $\mathbf{z}^{(j)}$ of Q are given by

$$z_i^{(j)} = 0 \qquad (1 \leq i \leq j),$$

$$z_i^{(j)} = (n-1)/2n(n-j) \qquad (j \leq i \leq n-1).$$

Thus the centroid $\mathbf{z}^*$ of Q has coordinates

$$z_i^* = (n-1)^{-1}\Sigma_{j<i}(n-1)/2n(n-j) = (1/2n)\Sigma_{j<i}(n-j)^{-1}, \tag{3.10}$$

which unlike (3.4) are strictly increasing in i.

We conclude by pointing out two somewhat disquieting features of the solution obtained above. The first [see (3.3)] is that the chance-event outcome ranked most probable is assigned a probability greater than 1/2, though only slightly so for large n. The second [see (3.2)] is the size of $F_{\min}$ . . . nearly 1/2 for large n. But perhaps this should not be surprising in view of the extremely conservative nature of minimax decision criteria.

(Paper 72B4–277)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 72B, No. 4, October–December 1968

On the Signs of the ν-Derivatives of the Modified Bessel Functions $I_\nu(x)$ and $K_\nu(x)$*

D. O. Reudink**

Bell Telephone Laboratories, Incorporated, Holmdel, New Jersey 07733

(September 17, 1968)

It is proved that $\partial I_\nu(x)/\partial\nu$ is negative and $\partial K_\nu(x)/\partial\nu$ is positive when $x>0$ and $\nu>0$.

Key Words: Bessel functions; ν-derivatives.

1. Introduction

The properties of the modified Bessel functions of the third kind have been investigated for many years. In a private communication, F. W. J. Olver pointed out that $\partial K_\nu(x)/\partial\nu>0$ when $x>0$ and $\nu>0$. This follows immediately from the integral representation for $K_\nu(x)$

$$K_\nu(x)=\int_0^\infty e^{-x\cosh t}\cosh(\nu t)dt. \tag{1}$$

It is the purpose of this paper to show that a similar fundamental property holds for $I_\nu(x)$, namely, $\partial I_\nu(x)/\partial\nu<0$ when $x>0$ and $\nu>0$.

The author wishes to acknowledge his discussions with F. Oberhettinger concerning this result.

2. Proof

Consider the following known integral[1]

$$I_\nu(x)K_\nu(x)=\frac{2}{\pi^2}\int_0^\infty\frac{\lambda\sinh(\pi\lambda)}{\lambda^2+\nu^2}K_{i\lambda}^2(x)d\lambda\ \text{Re}\ \nu>0. \tag{2}$$

Now differentiate the above formula with respect to ν to obtain an expression for $\partial I_\nu(x)/\partial\nu$,

$$\frac{\partial I_\nu(x)}{\partial\nu}=-\frac{1}{K_\nu(x)}\left[I_\nu(x)\frac{\partial K_\nu(x)}{\partial\nu}+\frac{4\nu}{\pi^2}\int_0^\infty\frac{\lambda\sinh(\pi\lambda)}{(\lambda^2+\nu^2)^2}K_{i\lambda}^2(x)d\lambda\right]. \tag{3}$$

It is easily seen from (1) that $K_{i\lambda}(x)$ is real when λ is real and $x>0$; hence for $\nu>0$, the integral in (3) is nonnegative. Therefore, since $K_\nu(x)$, $I_\nu(x)$, and $\partial K_\nu(x)/\partial\nu$ are positive for $\nu>0$ and $x>0$, it follows immediately that

$$\frac{\partial I_\nu(x)}{\partial\nu}<0 \qquad x>0, \qquad \nu>0. \tag{4}$$

---

*An invited paper.

**Present address: Bell Telephone Laboratories, Inc., R-133 Crawford Hill Lab., Box 400, Holmdel, New Jersey 07733.

[1] This integral is given in [1, p. 176, No. 8] for ν a positive integer, but is has been shown [2] that the expression is valid for Re $\nu>0$. Figures in brackets indicate the literature references at the end of this paper.

3. References

[1] Erdélyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G., Tables of Integral Transforms, Vol. 2 (McGraw-Hill Book Co., New York, 1954).
[2] Oberhettinger, F., On the Diffraction and Reflection of Waves and Pulses by Wedges and Corners, J. Res. NBS 61, No. 5, 343–365 (1958) RP2906.

(Paper 72b4–278)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 72B, No. 4, October–December 1968

Comparison of Finite-Difference Computations of Natural Convection

K. E. Torrance*

Institute for Applied Technology, National Bureau of Standards, Washington, D.C. 20234

(August 16, 1968)

Five numerical methods were compared for calculating two-dimensional, transient natural convection in an enclosure. Both implicit and explicit procedures were considered. Requirements for numerical stability were derived from analysis and experience, and when satisfied, the calculated flows for all methods were found to be similar. Consideration was also given to the accuracy and (energy and vorticity) conservation of the methods. One method was found to be conservative and stable without a restriction on the spatial mesh increment. This method can be successfully applied to nonlinear flows, but care must be exercised due to the presence of truncation errors which introduce false transport mechanisms.

Key Words: Enclosures; finite-difference; fluid flow; natural convection; numerical.

1. Introduction

1.1. General

The purpose of this paper is to compare the practical stability, truncation errors, and conservation properties of five finite-difference procedures for solving a problem in natural convection. Three of these methods are from the current literature [2, 6, 20][1] and two are developed here. The ability of the various methods to produce physically meaningful solutions is examined; and solutions calculated with the various methods are compared.

The physical problem chosen for study is the natural convection flow induced in a vertical circular cylinder by a small hot spot centrally located on the floor (see fig. 1). Transient and steady laminar flows are considered in two-dimensional, axisymmetric cylindrical coordinates. Such a fluid motion is described by three simultaneous partial differential equations [4, 10]: (a) an equation relating stream function to vorticity, eq (2); (b) a time-dependent equation for vorticity, eq (3); and (c) a time-dependent equation for temperature, eq (4). The first equation is of elliptic type, and the last two are of parabolic type.

1.2. Numerical Methods Tested

This section provides a brief introduction to the five numerical methods tested. All methods employ successive over-relaxation [16, see chapter 11] for solving the finite-difference approximation of the elliptic equation, and no difficulties are encountered. The parabolic equations, on the other hand, contain nonlinear terms with first-order derivatives which express the influence of convection (hereafter called the convection terms). These terms introduce serious problems of stability and conservation into the finite-difference scheme. The numerical methods differ only in

*Research Associate from the Factory Mutual Engineering Corporation, Norwood, Massachusetts. Present address: Department of Thermal Engineering, Upson Hall, Cornell University, Ithaca, N.Y. 14850.

[1] Figures in brackets indicate the literature references at the end of this paper.

321-453 OL-69—3

their treatment of the parabolic equations, and these differences are highlighted in table 1. Methods IV and V are new and are modifications of methods II and I, respectively.

TABLE 1. *Numerical methods applied to parabolic equations for temperature and vorticity*

Checks mean that the numerical methods possess the indicated properties.

Method	Reference	Explicit or implicit	Difference form for convection terms	Order of truncation error	No time step restriction	No spatial mesh size restriction	No false diffusion and convection	Satisfies conservation
I	Barakat and Clark [2]	E	2 pt forward or backward	h		✓		
II	Wilkes and Churchill [20]	I	3 pt central	h^2			✓	
III	Fromm [6]	E	3 pt central	h^2			✓	✓
IV	Present study	I	3 pt central	h^2			✓	✓
V	Present study	E	3 pt noncentral forward or backward	h		✓		✓

An inspection of table 1 reveals that no one method is conservative, free of false transport, and stable without a restriction on the spatial mesh size. For conservation reasons, methods III–V are to be preferred over methods I and II. Of these, methods III and IV require a reduction in spatial mesh size with increasing flow velocities to achieve stability, while method V does not. This mesh size restriction leads to prohibitively large demands on computer time and storage. The stability of method V, together with its conservation, are essential for calculating flows where the convection terms are important, as in the present problem. The introduction of false diffusion and convection, however, means that the calculated flows must be interpreted with care. In general, this false transport is important only in those flow regions where methods III and IV were found to break down.

Since the completion of this work, two additional differencing schemes have come to the author's attention [15, 1]. The first of these [15] was independently developed for steady flows and employs a differencing of the convection terms identical to method V. The second method [1] has been employed for meteorological problems and was developed for the vorticity equation. This method conserves vorticity, mean kinetic energy, and mean square vorticity. The convection terms are differenced to h^2 accuracy, as with methods II–IV in table 1, but conservation of mean square vorticity should lead to improved stability characteristics.

A comparison of the methods listed in table 1 is presented in the following sections. The physical problem is posed in section 2, some features and properties of the numerical methods are discussed in section 3, and comparative physical results are presented in section 4.

2. Mathematical Description of the Physical Problem

Consider the motion of a viscous fluid within a vertical circular cylinder of height a and radius b (see fig. 1). Erect a cylindrical coordinate system (x, r) with origin at the center of the base. The flow is assumed to be axisymmetric with no variations in the azimuthal direction. The fluid is initially motionless and at a uniform temperature T_0. The enclosure walls are also at this temperature, except for a small centrally located circular spot on the base of radius c which is at a temperature $T_h > T_0$. The temperature difference $T_h - T_0$ initiates and sustains the natural convection flow within the enclosure.

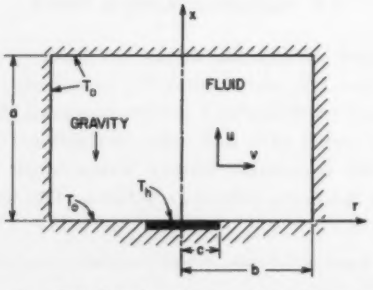


FIGURE 1. *Cylindrical enclosure and coordinate system.*

The Boussinesq approximation is used [4, see p. 16]; in this density (ρ) is assumed constant except for the generation of buoyancy forces. Other fluid properties are taken as constant: kinematic viscosity (ν), thermal diffusivity (κ), and volume expansion coefficient (β). We introduce the following dimensionless quantities: time, $\tau=(\kappa/a^2)t$; vertical and radial coordinates, $X=x/a$ and $R=r/a$; vertical and radial components of velocity, $U=(a/\kappa)u$ and $V=(a/\kappa)v$; and temperature $\Theta=(T-T_0)/(T_b-T_0)$.

The governing equations in dimensionless form are

$$U=\frac{1}{R}\frac{\partial\Psi}{\partial R},\ V=-\frac{1}{R}\frac{\partial\Psi}{\partial X}, \tag{1}$$

$$-\Omega=\frac{1}{R}\frac{\partial^2\Psi}{\partial X^2}+\frac{\partial}{\partial R}\left(\frac{1}{R}\frac{\partial\Psi}{\partial R}\right), \tag{2}$$

$$\frac{\partial\Omega}{\partial\tau}+\left[\frac{\partial(U\Omega)}{\partial X}+\frac{\partial(V\Omega)}{\partial R}\right]=-GrPr^2\frac{\partial\Theta}{\partial R}+Pr\left[\frac{\partial^2\Omega}{\partial X^2}+\frac{\partial}{\partial R}\left(\frac{1}{R}\frac{\partial(R\Omega)}{\partial R}\right)\right], \tag{3}$$

and

$$\frac{\partial\Theta}{\partial\tau}+\left[\frac{\partial(U\Theta)}{\partial X}+\frac{1}{R}\frac{\partial(RV\Theta)}{\partial R}\right]=\left[\frac{\partial^2\Theta}{\partial X^2}+\frac{1}{R}\frac{\partial}{\partial R}\left(R\frac{\partial\Theta}{\partial R}\right)\right]. \tag{4}$$

Equation (3) contains the Prandtl number $Pr=\nu/\kappa$ and the Grashof number $Gr=g\beta(T_h-T_0)a^3/\nu^2$. The latter number is used for natural convection flows and denotes the product of buoyancy and inertia forces divided by the square of the viscous force. The acceleration of gravity is denoted by g. The existence of a stream function Ψ is assumed such that the velocities are given by (1). The mass conservation equation div $\bar{U}=0$ is then automatically satisfied. The vorticity vector $\bar{\Omega}=$ curl $\bar{U}$ has only an azimuthal component, $\Omega=\partial V/\partial X-\partial U/\partial R$, given by (2). The conservation equations for vorticity and energy are (3) and (4) respectively. In these equations, bracketed terms on the left and right sides respectively denote the convective and diffusive transport terms. The term containing Gr in (3) represents the vorticity source due to buoyancy.

The convective terms in eqs (3) and (4) are in a "conservation" form [3, 6] suitable for methods III, IV, and V of table 1. Alternate "nonconservation" forms for these terms are utilized for methods I and II. The alternate equations are obtained by introducing a modified vorticity $\Omega'=(1/R)\Omega$ into (2) and (3), followed by subtraction of Ω' div $\bar{U}=0$ and Θ div $\bar{U}=0$ from (3) and (4) respectively. In terms of modified vorticity Ω', eqs (1) through (4) become

$$U=\frac{1}{R}\frac{\partial\Psi}{\partial R},\ V=-\frac{1}{R}\frac{\partial\Psi}{\partial X}, \tag{1'}$$

$$-R\Omega'=\frac{1}{R}\frac{\partial^2\Psi}{\partial X^2}+\frac{\partial}{\partial R}\left(\frac{1}{R}\frac{\partial\Psi}{\partial R}\right), \tag{2'}$$

$$\frac{\partial\Omega'}{\partial\tau}+U\frac{\partial\Omega'}{\partial X}+V\frac{\partial\Omega'}{\partial R}=-\frac{GrPr^2}{R}\frac{\partial\Theta}{\partial R}+Pr\left[\frac{\partial^2\Omega'}{\partial X^2}+\frac{1}{R}\frac{\partial}{\partial R}\left(\frac{1}{R}\frac{\partial(R^2\Omega')}{\partial R}\right)\right], \tag{3'}$$

and

$$\frac{\partial\Theta}{\partial\tau}+U\frac{\partial\Theta}{\partial X}+V\frac{\partial\Theta}{\partial R}=\frac{\partial^2\Theta}{\partial X^2}+\frac{1}{R}\frac{\partial}{\partial R}\left(R\frac{\partial\Theta}{\partial R}\right). \tag{4'}$$

Equations (1)–(4) and (1′)–(4′) are subject to initial conditions and boundary conditions. The initial conditions are:

$$\Omega=\Theta=0 \text{ for } \quad \tau<0,\ 0\leq X\leq 1, \text{ and } 0\leq R\leq R_b. \tag{5), (5'}$$

The boundary conditions[2] for $\tau \geq 0$ are:

$$\left.\begin{array}{lll} \Psi = \partial\Psi/\partial X = 0 & \text{for } X = 0, & 0 \leq R \leq R_b; \\ \Theta = 1 & \text{for } X = 0, & 0 \leq R < R_c; \\ \Theta = 0.5 & \text{for } X = 0, & R = R_c; \\ \Theta = 0 & \text{for } X = 0, & R_c < R \leq R_b; \\ \Psi = \partial\Psi/\partial X = \Theta = 0 & \text{for } X = 1, & \text{all } R; \\ \Psi = \Omega = \partial\Theta/\partial R = 0 & \text{for } R = 0, & \text{all } X; \text{ and} \\ \Psi = \partial\Psi/\partial R = \Theta = 0 & \text{for } R = R_b, & \text{all } X. \end{array}\right\} \quad (6), (6')$$

For equations (1′)–(4′) replace $\Omega = 0$ in (5) by $\Omega' = 0$, and $\Omega = 0$ ($R = 0$, all X) in (6) by the dual conditions $V = \partial U/\partial R = 0$.[3]

The boundary conditions introduce two geometric parameters: the aspect ratio of the enclosure (radius/height) $R_b = b/a$ and the relative size of the heat source (heat source radius/enclosure height) $R_c = c/a$.

To determine Θ and Ω' along the centerline, special forms of the conservation equations are needed to avoid an indeterminate form as $R \to 0$. By incorporating boundary conditions at $R = 0$ and using L'Hospital's rule, eqs (4), (3′), and (4′) respectively reduce to

$$\frac{\partial\Theta}{\partial\tau} + \frac{\partial(U\Theta)}{\partial X} + \frac{\partial^2(RV\Theta)}{\partial R^2} = \frac{\partial^2\Theta}{\partial X^2} + 2\,\frac{\partial^2\Theta}{\partial R^2}, \quad (7)$$

$$\frac{\partial\Omega'}{\partial\tau} + U\,\frac{\partial\Omega'}{\partial X} = -GrPr^2\,\frac{\partial^2\Theta}{\partial R^2} + Pr\left[\frac{\partial^2\Omega'}{\partial X^2} + 4\,\frac{\partial^2\Omega'}{\partial R^2}\right], \quad (7'a)$$

and

$$\frac{\partial\Theta}{\partial\tau} + U\,\frac{\partial\Theta}{\partial X} = \frac{\partial^2\Theta}{\partial X^2} + 2\,\frac{\partial^2\Theta}{\partial R^2}. \quad (7'b)$$

For convenience, the velocities U and V are retained explicitly in the problem formulation. However, an examination of eqs (1)–(7) (or alternately, eqs (1′)–(7′)) reveals that the essential dependent variables are Ψ, Ω, and Θ (or Ψ, Ω', and Θ), while the independent variables are X, R, and τ. The parameters of the problem are R_b, R_c, Pr, and Gr. Throughout this study, the aspect ratio, relative heat source size, and Prandtl number are held fixed at $R_b = 1$, $R_c = 0.1$ and $Pr = 0.7$, respectively. The Grashof number Gr is assumed equal to 1×10^5, except in section 4.4, which considers larger values of this parameter.

3. Numerical Methods, Formulation and Some Properties

3.1. Grid System and Calculation Sequence

An approximate solution of eqs (1)–(7) or (1′)–(7′) will be obtained at a finite number of grid points having coordinates $X = i\Delta X$, $R = j\Delta R$, and at discrete times τ_n, where i, j, and n are integers. The grid spacings in the X and R directions are denoted by ΔX and ΔR. The symbol τ_n denotes the time level after the nth time step $\Delta\tau_n$. The values of Ψ, Ω, Θ, U, and V at each grid point should be thought of as average values over a small volume of fluid surrounding the point.

---

[2] Note that explicit boundary conditions for vorticity on the solid boundaries are not available. This causes no difficulty with the explicit solution methods and only slight difficulty with the implicit methods.

[3] The centerline boundary condition $\Omega = R\Omega' = 0$ is satisfied for all finite values of Ω'. Although the vorticity Ω is zero, the modified vorticity Ω' is not, in general, zero. Barakat and Clark [2] used the modified vorticity Ω' and assumed $\Omega' = 0$ along the centerline. This is incorrect, and may have contributed to the oscillations in stream function (near the centerline) and heat transfer observed in their calculations. The vorticity Ω' on the centerline can be evaluated with the aid of eq (7′a).

All of the numerical methods listed in table 1 advance the fields of Ψ, Ω, and Θ across a time step in the same sequence. All quantities are assumed known at a time τ_n. The sequence for advancement from time τ_n to the new level $\tau_{n+1}=\tau_n+\Delta\tau$ is as follows:

1. Temperature (Θ) at all interior grid points is advanced with suitable difference approximations of (4) and (7) or (4′) and (7′b). This is discussed in section 3.2.

2. Vorticity (Ω or Ω') at all interior grid points is similarly advanced with approximations of (3) or (3′) and (7′a). This is discussed in section 3.2.

3. Stream function (Ψ) at all interior grid points is brought up to date with the new vorticity field by using a difference approximation of (2) or (2′). This is discussed in section 3.3.

4. The vorticity on the solid boundaries is determined from the new stream function field, as are the velocities U and V. This is discussed in section 3.3.

5. The field of mesh points is scanned in order to determine the time step $\Delta\tau$ for the next time advancement. This is limited by practical stability requirements and is discussed in section 3.4.

3.2. Parabolic Equations

a. Differencing Schemes

The approximations of the equations for temperature and vorticity (for steps 1 and 2 above) are described in conventional notation in table 2. In this section we comment on the distinguishing features and some of the properties of these schemes.

TABLE 2. *Finite-difference approximations of the parabolic equations for temperature and vorticity*

Illustrated with the simplified temperature equations, $\frac{\partial\Theta}{\partial\tau}+U\frac{\partial\Theta}{\partial X}-\frac{\partial^2\Theta}{\partial X^2}=0$ (for methods I and II) and $\frac{\partial\Theta}{\partial\tau}+\frac{\partial(U\Theta)}{\partial X}-\frac{\partial^2\Theta}{\partial X^2}=0$ (for methods III–V).

Method		$\frac{\partial\Theta}{\partial\tau}$	$U\frac{\partial\Theta}{\partial X}$ or $\frac{\partial(U\Theta)}{\partial X}$	$\frac{\partial^2\Theta}{\partial X^2}$
I		$\frac{\Theta^{n+1}_{i,j}-\Theta^n_{i,j}}{\Delta\tau}+0(\Delta\tau)$	$U\frac{\partial\Theta}{\partial X}=\begin{cases}U^n_{i,j}\frac{\Theta^n_{i,j}-\Theta^n_{i-1,j}}{\Delta X}, & (U^n_{i,j}>0)\\ U^n_{i,j}\frac{\Theta^n_{i+1,j}-\Theta^n_{i,j}}{\Delta X}, & (U^n_{i,j}<0)\end{cases}+\Delta X\frac{\lvert U\rvert}{2}\frac{\partial^2\Theta}{\partial X^2}+0(\Delta X)^2$	$\frac{\Theta^n_{i+1,j}-2\Theta^n_{i,j}+\Theta^n_{i-1,j}}{(\Delta X)^2}+0(\Delta X)^2$.
II	a	$\frac{\Theta^{n+1/2}_{i,j}-\Theta^n_{i,j}}{\Delta\tau/2}+0(\Delta\tau)$	$U\frac{\partial\Theta}{\partial X}=U^n_{i,j}\frac{\Theta^n_{i+1,j}-\Theta^n_{i-1,j}}{2\Delta X}+0(\Delta X)^2$	$\frac{\Theta^n_{i+1,j}-2\Theta^n_{i,j}+\Theta^n_{i-1,j}}{(\Delta X)^2}+0(\Delta X)^2$.
	b	$\frac{\Theta^{n+1}_{i,j}-\Theta^{n+1/2}_{i,j}}{\Delta\tau/2}+0(\Delta\tau)$	$U\frac{\partial\Theta}{\partial X}=U^n_{i,j}\frac{\Theta^{n+1}_{i+1,j}-\Theta^{n+1}_{i-1,j}}{2\Delta X}+0(\Delta X)^2+0\left(\frac{\Delta\tau}{\Delta X}\right)$	$\frac{\Theta^{n+1}_{i+1,j}-2\Theta^{n+1}_{i,j}+\Theta^{n+1}_{i-1,j}}{(\Delta X)^2}+0(\Delta X)^2+0\left(\frac{\Delta\tau}{\Delta X}\right)$.
III		$\frac{\Theta^{n+1}_{i,j}-\Theta^{n-1}_{i,j}}{2\Delta\tau}+0(\Delta\tau)^2$	$\frac{\partial(U\Theta)}{\partial X}=\frac{U^n_{i+1,j}\Theta^n_{i+1,j}-U^n_{i-1,j}\Theta^n_{i-1,j}}{2\Delta X}+0(\Delta X)^2$	$\frac{\Theta^n_{i+1,j}-\Theta^{n-1}_{i,j}+\Theta^n_{i-1,j}}{(\Delta X)^2}+0(\Delta X)^2+0\left(\frac{\Delta\tau}{\Delta X}\right)^2$.
IV	a	Same as method IIa.	$\frac{\partial(U\Theta)}{\partial X}=\frac{U^n_{i+1,j}\Theta^n_{i+1,j}-U^n_{i-1,j}\Theta^n_{i-1,j}}{2\Delta X}+0(\Delta X)^2$	Same as method IIa.
	b	Same as method IIb.	$\frac{\partial(U\Theta)}{\partial X}=\frac{U^n_{i+1,j}\Theta^{n+1}_{i+1,j}-U^n_{i-1,j}\Theta^{n+1}_{i-1,j}}{2\Delta X}+0(\Delta X)^2+0\left(\frac{\Delta\tau}{\Delta X}\right)$	Same as method IIb.
V		Same as method I.	$\frac{\partial(U\Theta)}{\partial X}=\begin{cases}\frac{\bar U^n_{i,j}\Theta^n_{i,j}-\bar U^n_{i-1,j}\Theta^n_{i-1,j}}{\Delta X}, & (\bar U^n_{i,j},\ \bar U^n_{i-1,j}>0)\\ \frac{\bar U^n_{i,j}\Theta^n_{i+1,j}-\bar U^n_{i-1,j}\Theta^n_{i,j}}{\Delta X}, & (\bar U^n_{i,j},\ \bar U^n_{i-1,j}<0)\end{cases}+0(\Delta X)+0(\Delta X)^2$*	Same as method I.

*$\bar U_{m,n}=(U_{m+1,n}+U_{m,n})/2$.

Application of the various numerical methods to the parabolic equations is illustrated with the one-space dimension temperature equation

$$\frac{\partial \Theta}{\partial \tau}+\frac{\partial (U\Theta)}{\partial X}-\frac{\partial^2 \Theta}{\partial X^2}=0 \tag{8}$$

in place of eqs (3), (4), and (7); and

$$\frac{\partial \Theta}{\partial \tau}+U\frac{\partial \Theta}{\partial X}-\frac{\partial^2 \Theta}{\partial X^2}=0 \tag{8'}$$

in place of eqs (3′), (4′), and (7′). Equation (8) is employed for methods III-V and eq (8′) for methods I and II (see table 2). It is understood that the discussion and differencing of each term of these equations applies to similar terms in the two-space dimension temperature and vorticity equations.

Methods I, III, and V are explicit procedures for calculating quantities at time $n+1$ in terms of their values at times n and $n-1$. Methods II and IV are two step implicit procedures (see [12, 16, 20]), going from time n to $n+1/2$ in step a, and from $n+1/2$ to $n+1$ in step b. During the first half time step (or second half time step), the X and R space derivatives are respectively approximated at time levels n and $n+1/2$ (or $n+1$ and $n+1/2$). Clearly, the time steps a and b are implicit in the R and X directions, respectively. This is called an alternating direction, implicit scheme and leads to a tridiagonal matrix of unknown temperatures or vorticities for each R-row or X-column, as the case may be. These matrices are readily inverted by a simple algorithm.

Column two of table 2 lists the approximation of the derivative $\partial\Theta/\partial\tau$. A central time difference is used with method III, and forward time differences with all others. The last column of the table lists the approximation of the diffusion terms, $\partial^2\Theta/\partial X^2$. Three point space differences are used for all methods.

The approximations of the convection terms $U\partial\Theta/\partial X$ or $\partial(U\Theta)/\partial X$ are listed in column three. At any grid point the velocities are evaluated at time n and are treated as constants over the time step. With method I, the convection terms are approximated with two-point forward or backward differences as the coefficient velocity U is positive or negative, respectively. Methods II–IV all employ three-point central differences for these terms. Method V employs a modified form of forward or backward differences in which the mean velocity $\bar{U}_{n,m}$ in the numerator is multiplied by $\Theta_{n,m}$ or $\Theta_{n+1,m}$ as $\bar{U}_{n,m}$ is positive or negative, respectively. If the two mean velocity coefficients are of different sign, one term from the numerator of each of the two approximations shown is required.

Two points about the vorticity equation should be noted. First, this equation contains a buoyancy source term which is not listed in table 2. The term contains $\partial\Theta/\partial R$, which is approximated with three-point central differences. Method III evaluates the term at time n, all other methods evaluate it at time $n+1$. Second, when the implicit methods II and IV are applied to the vorticity equation, it is necessary to temporarily assume the vorticity distribution on the solid walls at times $n+1$ and $n+1/2$ equal to that at time n. Thus, the wall vorticity is out of step with the advancement of the interior vorticity field. The error introduced by this assumption decreases with both small time steps and the approach to steady state.

The rest of this subsection presents brief comments on the stability and conservation properties of the parabolic difference equations. Additional discussion is provided in sections 3.4 and 4.3, respectively.

The stability of the various numerical methods is principally due to some key approximations shown in table 2. For methods I and V, it is the use of forward or backward differencing for the convection terms [2, 7, 14 (see p. 194)]. For method III, it is the duFort-Frankel differencing [5] of the diffusion terms $\partial^2\Theta/\partial X^2$. For methods II and IV, it is the alternating direction, implicit nature of the scheme (due to Peaceman-Rachford [12, 16 (see p. 366), 20]).

A lengthy but straightforward study of methods I–V shows that conservation is satisfied for III–V, but not for I and II. The elements of such a study are presented in [3, 6]. Conservation of energy or vorticity within the grid system exists if the difference equations for temperature or vorticity are summed over all interior grid points and no spurious sources or sinks of these quan-

tities are found. The convection and diffusion terms cancel in pairs. The net energy or vorticity transport from the wall mesh points into the enclosure just balances the net increase of energy or vorticity within the mesh system. Heat transfer rates to and from the enclosure can thus be determined by calculating the transfer from wall mesh points. If conservation is satisfied, there is no net increase of energy within the enclosure at steady state, and the total heat added to the enclosure balances the heat removed.

b. Truncation Errors

The truncation errors shown in table 2 are worthy of some discussion. These estimates are obtained by Taylor series analysis as described in references [5, 11, 13]. Methods II–IV are most easily discussed and are treated in the first paragraph. Methods I and V are more complicated, and are discussed in the second and third paragraphs, respectively.

For methods II–IV, all space derivatives are approximated to $(\Delta X)^2$ accuracy. The use of multiple time levels in conjunction with these space derivatives, however, introduces additional truncation errors of $0(\Delta\tau/\Delta X)$ for methods II and IV and of $0(\Delta\tau/\Delta X)^2$ for method III. Substitution of the difference approximations listed in table 2 into eqs (8) or (8′), as appropriate, introduces truncation errors of $0(\Delta\tau)+0(\Delta\tau/\Delta X)+0(\Delta X)^2$ for methods II and IV and of $0(\Delta\tau)^2+0(\Delta\tau/\Delta X)^2+0(\Delta X)^2$ for method III. The coefficients of these terms involve derivatives of higher order than those of the basic eqs (8) or (8′). Stability requires that $\Delta\tau \ll \Delta X$, and when ΔX is small, these truncation errors are usually neglected.

With method I, the $0(\Delta X)$ approximation of the convection term $U\partial\Theta/\partial X$ introduces a large truncation error. A Taylor series expansion for $\Theta_{i-1,j}$ about $\Theta_{i,j}$ can be rearranged to solve for $(\partial\Theta/\partial X)_{i,j}$

$$\left(\frac{\partial\Theta}{\partial X}\right)_{i,j}=\frac{\Theta_{i,j}-\Theta_{i-1,j}}{\Delta X}+\left\{\frac{\Delta X}{2}\left(\frac{\partial^2\Theta}{\partial X^2}\right)_{i,j}-\frac{(\Delta X)^2}{6}\left(\frac{\partial^3\Theta}{\partial X^3}\right)_{i,j}+\cdots\right\}. \tag{9}$$

The first term on the right side is a backward difference. The term in braces is the truncation error. Multiplication of eq (9) by $U_{i,j}(U_{i,j}>0)$ leads to the form shown in the table. The solution of the difference equation (i.e., the approximation of (8′)) is equivalent to a solution of the differential equation

$$\frac{\partial\Theta}{\partial\tau}+U\frac{\partial\Theta}{\partial X}-\left(1+\Delta X\frac{|U|}{2}\right)\frac{\partial^2\Theta}{\partial X^2}=0(\Delta\tau)+0(\Delta X)^2. \tag{10}$$

The coefficients of the terms $0(\Delta\tau)$ and $0(\Delta X)^2$ involve derivatives of higher order than appear in (8′), and for reasons noted in the preceding paragraph, these truncation errors are usually neglected. The differencing of the convection terms in equation (8′) introduces an additional, or false, heat diffusivity $\Delta X\frac{|U|}{2}$. Only for small $\Delta X\frac{|U|}{2}$ does the difference approximation approach the differential equation (8′). For large $\Delta X\frac{|U|}{2}$, care must be exercised because a false heat diffusivity is introduced (or a false viscosity in the case of the vorticity equation).[4]

Turning now to method V, the analysis is similar to that presented for method I above. The truncation error of $0(\Delta X)$ in the convection term $\partial(U\Theta)/\partial X$ is more complicated, however. It can be shown that the solution of the difference equation (i.e., the approximation of (8)) is equivalent to a solution of the differential equation

$$\frac{\partial\Theta}{\partial\tau}+\left\{\frac{\partial(U\Theta)}{\partial X}-\frac{\Delta X}{2}\frac{\partial U}{\partial X}\frac{\partial\Theta}{\partial X}\right\}-\left\{1+\Delta X\frac{\bar{U}_{i,j}}{2}\right\}\frac{\partial^2\Theta}{\partial X^2}=0(\Delta\tau)+0(\Delta X)^2 \tag{11}$$

[4] This problem cannot be overcome by eliminating $\partial^2\Theta/\partial X^2$ from (9) with, for example, three-point central differences. If this is done, eq (9) becomes a three-point central difference approximation of $\partial\Theta/\partial X$, with its associated stability problems.

when $\bar{U}_{i,j}$, $\bar{U}_{i-1,j} > 0$. Again, the terms on the right side are usually neglected. There is now a false heat diffusion (or a false viscosity) as well as a false convection. The latter is a subtraction from the real convection and is small whenever $\frac{\Delta X}{2}(\partial U/\partial X)(\partial \Theta/\partial X)$ is small. When this last term and $\Delta X \frac{U}{2}$ are small, the difference approximation approaches the differential equation (8). When these terms are large, care must be exercised.

3.3. Elliptic Equation, Velocities, and Wall Vorticity

The computations described in steps 3 and 4 of section 3.1 are identical for all five numerical methods, and are integrated and presented in this section. Equations for the vorticity Ω (used for methods III–V) are considered here. Appropriate equations in terms of modified vorticity Ω' (for methods I and II) are obtained by the substitution $\Omega = R\Omega'$. Step 3 is considered first.

The elliptic equation (2) relates stream function to vorticity at any given time. The new vorticities $\Omega_{i,j}^{n+1}$ are introduced and the space derivatives are approximated by three-point central differences with truncation errors of $0(\Delta X)^2$ and $0(\Delta R)^2$. An iterative technique known as successive over-relaxation [21] was employed to obtain the new stream function field. No problems of stability or conservation are encountered. Thus, if $\Psi_{i,j}^{(s)}$ denotes the approximate stream function at a point interior to the solid boundaries and centerline after s iterations, a further approximation $\Psi_{i,j}^{(s+1)}$ is obtained from

$$\Psi_{i,j}^{(s+1)} = (1-\omega)\Psi_{i,j}^{(s)} + \frac{\omega}{\frac{2}{(\Delta X)^2} + \frac{1}{(\Delta R)^2}\left(\frac{j}{j+\frac{1}{2}} + \frac{j}{j-\frac{1}{2}}\right)} \left\{ j\Delta R\Omega_{i,j}^{n+1} + \frac{1}{(\Delta X)^2}\left[\Psi_{i+1,j}^{(s)} + \Psi_{i-1,j}^{(s+1)}\right] + \frac{1}{(\Delta R)^2}\left[\frac{j}{j+\frac{1}{2}}\Psi_{i,j+1}^{(s)} + \frac{j}{j-\frac{1}{2}}\Psi_{i,j-1}^{(s+1)}\right]\right\}. \quad (12)$$

It is understood that all values of Ψ in this equation pertain to time $n+1$. The optimum value of the relaxation parameter ω was calculated for a given system of grid points (see eqs (1.9) and (4.6) of reference [21]).[5]

Successive sweeps of eq (12) over the field of mesh points were alternated in the vertical and radial directions. Iteration was terminated when

$$\frac{\max_{i,j} |\Psi_{i,j}^{(s+1)} - \Psi_{i,j}^{(s)}|}{\max_{i,j} |\Psi_{i,j}^{(s+1)}|} < 0.0001. \quad (13)$$

The number 0.0001 was arrived at by experimentation, noting that more stringent criteria led to no essential difference in the solution of test problems. The numerator on the left side of (13) is the maximum change in stream function occurring in the field of grid points as a result of one approximation step. The denominator is the maximum value of stream function in the field.

In practice, the most time consuming part of the whole calculation was the iteration of eq (12) at each time step. One iterative sweep took almost as long as the entire computation of the new temperature or vorticity field described in the previous section. The following procedures resulted in a considerable savings in computer time. A maximum number of iterations, s_{max}, was permitted. For 100, 400, and 1000 mesh points, s_{max} was 20, 25, and 35, respectively. The convergence criterion (13) was generally achieved within s_{max} iterations after the initial flow transient was over. The stream function field was then updated less frequently, typically every $1 + 0.25\,(s_{max} - s_{convergence})$ time steps.

[5] An independent determination of the optimum ω was made with the aid of a computer. Equation (12) was solved for various ω starting with a given Ω field. The value of ω leading to fastest convergence was sufficiently close to the calculated value so that the latter was used.

The computations described in step 4 (sec. 3.1) are discussed in the rest of this section. The numerical approximations up to this point, together with the boundary conditions (6), provide the entire fields of Ψ and Θ, and the field of Ω interior to the solid boundaries, all at time $n+1$. Still to be calculated at time $n+1$ are the unknown values of vorticity Ω on the solid boundaries [6] and the velocity fields U and V. The new stream function field is used for this purpose.

The vorticity on the solid boundaries is obtained by first reducing eq (2) to $\Omega = -(1/R)\partial^2\Psi/\partial X^2$ for $X = 0$ or $X = 1$, and to $\Omega = -(1/R_b)\partial^2\Psi/\partial R^2$ for $R = R_b$. The second order derivatives are approximated by expanding Ψ in a Taylor series about the wall. Expansions for the two grid points nearest the wall, together with the boundary conditions on Ψ, yield wall vorticity approximations such as the following, which applies when $X = 0$:

$$\Omega_{0,j}^{n+1} = -\frac{8\Psi_{1,j} - \Psi_{2,j}}{2j\Delta R(\Delta X)^2} + 0(\Delta X)^2. \tag{14}$$

Finally, the fields of U and V are calculated with three point central difference approximations of eqs (1). These approximations (of $0(\Delta X)^2$ and $0(\Delta R)^2$) are not presented here, but it is readily verified that they automatically satisfy a central difference approximation of the continuity equation, $\operatorname{div} \vec{U} = 0$. The conservation of mass within the grid system is thereby established.

3.4. Practical Stability Considerations for the Parabolic Equations

At this point, the fields of Ψ, Ω, Θ, U, and V are current at time $n+1$ for all methods. Further integration of the parabolic temperature and vorticity equations in time and space requires consideration of practical stability. This is step 5 of section 3.1. Practical stability imposes a restriction on the size of the time step $\Delta\tau$ for methods I and V (eqs (19) and (20), respectively) and restrictions on $\Delta\tau$, ΔX, and ΔR for methods II–IV (eqs (15)). The size of $\Delta\tau$ is calculated from the appropriate equations (assuming for methods II–IV that the restrictions on ΔX and ΔR are satisfied). The fields of Ψ, Ω, Θ, U, and V are then advanced across this time step by repeating the whole cycle described in sections 3.2 and 3.3.

A complete analysis of the nonlinear equations to determine the exact form of the stability requirements is not always possible. Such is the case for methods II–IV. The linearized stability analysis of von Neumann [11] has been applied to methods II [19] and III [6] and leads only to a restriction on $\Delta\tau$ for method III. Numerical experience with methods II–IV for a limited range of flows suggests that the following *empirical* restrictions be applied at each grid point for integration of the temperature and vorticity equations:

$$\Delta X \leq \frac{8}{|U_{i,j}|},\ \Delta R \leq \frac{8}{|V_{i,j}|}, \tag{15a}$$

$$\Delta\tau \leq \left[\frac{2}{(\Delta X)^2} + \frac{2}{(\Delta R)^2}\right]^{-1}, \tag{15b}$$

and

$$\Delta\tau \leq \left[\frac{|U_{i,j}|}{\Delta X} + \frac{|V_{i,j}|}{\Delta R}\right]^{-1}. \tag{15c}$$

Note the restriction on spatial mesh size (15a), which presumably results from the use of three point central differences for the three methods. Within limited ranges these restrictions have led to stability, but they cannot be regarded as general. In particular, during flow transients with large initial changes, time steps smaller than suggested by (15b) or (15c) were required.

[6] This delay in calculating the wall vorticities causes no difficulty with the explicit methods I, III, and V. For the implicit methods II and IV, it was necessary in section 3.2a to temporarily assume that the wall vorticities at times $n+1$ and $n+1/2$ were equal to their values at time n.

Analytical stability criteria are possible for methods I and V. The analysis is applicable to explicit, nonlinear, two time level differencing schemes and is discussed in reference [2]. This paragraph is based on that analysis. To be specific, stability of the integration of the temperature eqs (4) or (4′) is considered here; that for vorticity (3) or (3′) is similar. In methods I and V, the unknown temperature $\Theta_{i,j}^{n+1}$ can be written as an explicit linear combination of computed values at time n:

$$\Theta_{i,j}^{n+1} = a_1\Theta_{i+1,j}^{n} + a_2\Theta_{i-1,j}^{n} + a_3\Theta_{i,j}^{n} + a_4\Theta_{i,j+1}^{n} + a_5\Theta_{i,j-1}^{n} \tag{16}$$

where the a_k denote coefficients which vary in time, but which are constant over a time step. The essential feature of the method is to require that the norm of the matrix of coefficients a_k at all times is bounded by unity. The row norm is used, and leads to

$$\max_{i,j}\left\{\sum_{k=1}^{5}|a_k|\right\} \leq 1 \tag{17}$$

where $\max_{i,j}$ denotes the maximum value of the quantity in braces for all i, j in the grid system. Stability in the sense of Lax and Richtmyer [9, 14 (see p. 44)] exists if inequality (17) is satisfied. For methods I and V, satisfying (17) is equivalent to requiring that all the coefficients a_k are positive. This follows because for methods I and V the coefficient a_3 has the special form:

$$a_3 = 1 - \{a_1 + a_2 + a_4 + a_5\} \tag{18}$$

which can be readily verified. The conditions under which the coefficients a_k are positive for methods I and V are discussed next.

For method I, the coefficients a_k ($k = 1, 2, 4, 5$) are always positive, whereas a_3 can be made positive by restricting the size of the time step $\Delta\tau$. For $Pr \leq 1$, the most severe restrictions on $\Delta\tau$ are:

$$\Delta\tau \leq \left[\frac{|U_{i,j}|}{\Delta X} + \frac{|V_{i,j}|}{\Delta R} + \frac{2}{(\Delta X)^2} + \frac{2}{(\Delta R)^2}\right]^{-1}, \tag{19a}$$

$$\Delta\tau \leq \left[\frac{|U_{i,0}|}{\Delta X} + \frac{2}{(\Delta X)^2} + \frac{4}{(\Delta R)^2}\right]^{-1}, \tag{19b}$$

and

$$\Delta\tau \leq \left[\frac{|U_{i,0}|}{\Delta X} + \frac{2Pr}{(\Delta X)^2} + \frac{8Pr}{(\Delta R)^2}\right]^{-1}. \tag{19c}$$

These restrictions respectively follow by requiring that a_3 be positive for the finite-difference representations of eqs (4′), (7′b), and (7′a).

For method V (as for method I above), the coefficients a_k ($k = 1, 2, 4, 5$) are always positive. In order to make a_3 positive, however, several different restrictions on $\Delta\tau$ arise because method V employs forms for the convection terms that depend upon the sign of the mean velocities, as noted in table 2. For the case when the mean velocities $\bar{U}_{i,j}$, $\bar{U}_{i-1,j}$, $\bar{V}_{i,j}$ and $\bar{V}_{i,j-1}$ are positive[7] and $Pr \leq 1$, the greatest restrictions on $\Delta\tau$ result from the difference forms of the temperature equations (4) and (7):

$$\Delta\tau \leq \left[\frac{\bar{U}_{i,j}}{\Delta X} + \left(1 + \frac{1}{2j}\right)\frac{\bar{V}_{i,j}}{\Delta R} + \frac{2}{(\Delta X)^2} + \frac{2}{(\Delta R)^2}\right]^{-1} \tag{20a}$$

[7] The mean velocities are defined by

$$\bar{U}_{m,n} = \frac{U_{m+1,n} + U_{m,n}}{2} \text{ and } \bar{V}_{m,n} = \frac{V_{m,n+1} + V_{m,n}}{2}$$

and

$$\Delta\tau \leq \left[\frac{\bar{U}_{i,\theta}}{\Delta X}+4\,\frac{\bar{V}_{i,\theta}}{\Delta R}+\frac{2}{(\Delta X)^2}+\frac{4}{(\Delta R)^2}\right]^{-1}. \tag{20b}$$

Additional forms similar to (20) arise when the mean velocities are negative, or a combination of positive and negative.

Some observations on the application of the stability requirements can be noted. It is impractical to alter ΔX and ΔR according to eqs (15a) during a computation to achieve stability for methods II–IV. Accordingly, a cut and try procedure is employed. The time step, however, is adjusted during a computation with all methods, using equations (15b), (15c), (19), or (20), as appropriate. The field of mesh points is scanned in order to determine the largest allowable $\Delta\tau$ which will satisfy the stability requirements. Time steps smaller than this value can be used, of course, to provide a safety factor. In practice, 80 percent of the allowable $\Delta\tau$ was used with methods I–IV and 95 percent with method V.

4. Comparative Physical Results

Natural convection flows calculated with the five numerical methods are presented and compared in sections 4.1 through 4.3 for a Grashof number of $Gr = 1 \times 10^5$. The calculations employed a uniform grid spacing of $\Delta X = \Delta R = 0.05$, with a total of 121 mesh points. The calculation of physically meaningful flows at higher values of Gr, and some questions of convergence, are considered in section 4.4.

Flow patterns and temperature fields are illustrated in figures 2 through 6 with computer-drawn graphs displaying sets of streamlines and isotherms. The location of these streamlines and isotherms was determined by linear interpolation of the computed values at the mesh points. In each of the graphs, the centerline of the cylindrical enclosure is shown on the left. The abscissa is the radial coordinate R, and the ordinate is the axial coordinate X. The heat source on the floor is denoted with a thick line between $R = 0$ and $R = 0.1$.

4.1. Transient Streamline and Temperature Fields

All the numerical calculations were carried from the initial quiescent condition forward in time until steady state was achieved. The transient and steady-state flows calculated during this process are qualitatively similar for all five numerical methods. The purpose of this section is to discuss the physical nature of the flows, consequently, only the transient results for one method need be considered. Method V is considered representative, and the transient flow and temperature fields calculated with this method are illustrated in figures 2 and 3, respectively. These figures are each composed of four graphs, arranged in order of increasing time, τ.

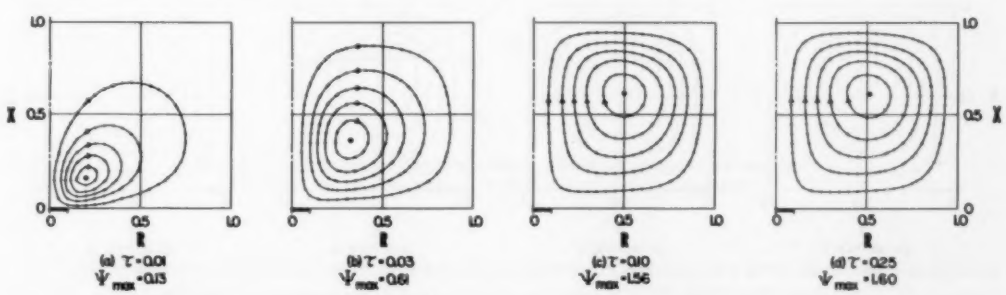


FIGURE 2. *Transient streamline fields calculated with method* V *at various times* (τ), $Gr = 1 \times 10^5$.

Streamline evolution is qualitatively similar for all five numerical methods. The walls and centerline correspond to $\Psi = 0$; the dot has value Ψ_{max}; the remaining streamlines correspond successively to Ψ values of 0.1, 0.3, 0.5, 0.7, and 0.9, of Ψ_{max}.

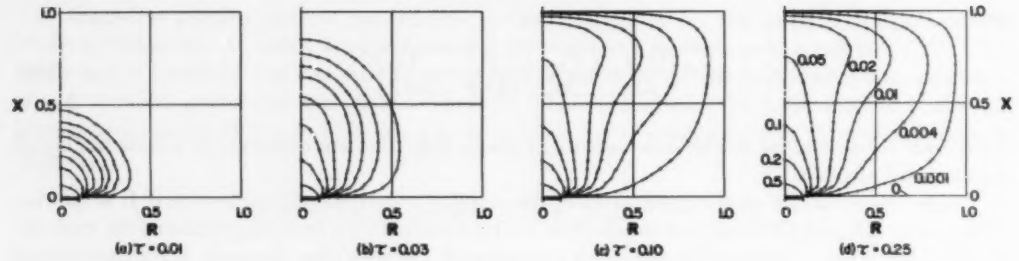


FIGURE 3. *Transient temperature fields calculated with method* V *at various times* (τ), Gr$= 1 \times 10^5$.

Temperature evolution is qualitatively similar for all five numerical methods. Temperature (Θ) is the curve parameter; the heat source (thick line on floor) is at $\Theta = 1$.

Immediately after the start of heating, a ring vortex of warm fluid forms near the origin (figs. 2a and 3a). This vortex rises and moves radially outward from the centerline (figs. 2b and 2c). A gradual development into the steady state flow and temperature fields then follows. The steady state fields reveal a ring vortex centered just above $X = 0.5$, $R = 0.5$ (fig. 2d) which is driven by the heated fluid which rises along the centerline from the floor to the ceiling (fig. 3d).

A sensitive indication of the approach to steady state was provided by the overall energy balance on the enclosure. When the heat transfer rates into and out of the enclosure were within 1 percent of an asymptotic value, steady conditions prevailed. Steady state was achieved by $\tau = 0.25$ with all methods.

4.2. Comparison of Steady-State Streamline and Temperature Fields

The qualitative similarity of flows calculated with the five methods was mentioned in the last section. The purpose of the present section is to illustrate this similarity in a quantitative way, by comparing steady state flows obtained with methods I–V. Such a comparison of the steady-state streamline and temperature fields is presented in figures 4 and 5, respectively. The sets of streamlines and isotherms shown have the same numerical values as in figures 2 and 3. Figures 4 and 5 each contain four individual graphs, which respectively pertain to numerical methods I, II, III, and V. Computed results from methods III and IV agreed to at least three significant figures, and graphs of the steady state streamline and temperature fields for the two methods are identical. Consequently, only the results of method III are shown. Figures 4d and 5d duplicate figures 2d and 3d respectively.

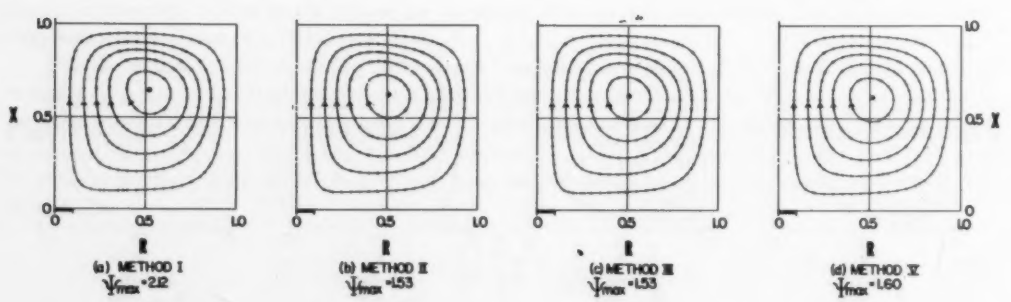


FIGURE 4. *Steady-state streamline fields for the various numerical methods*, Gr$= 1 \times 10^5$.

Results for methods III and IV are identical. The walls and centerline correspond to $\Psi = 0$; the dot has value Ψ_{max}; the remaining streamlines correspond successively to Ψ values of 0.1, 0.3, 0.5, 0.7, and 0.9, of Ψ_{max}.

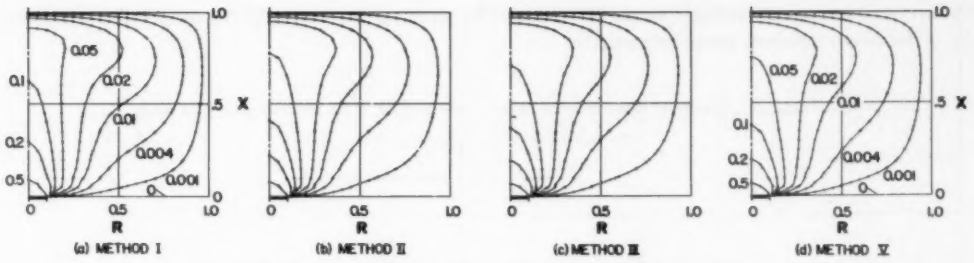


FIGURE 5. *Steady state temperature fields for the various numerical methods,* $Gr = 1 \times 10^6$.
Results for methods III and IV are identical. Temperature (Θ) is the curve parameter; the heat source (thick line on floor) is at $\Theta = 1$.

A comparison of figures 4a–4d or 5a–5d reveals a striking similarity in the streamline and temperature fields computed with the various numerical methods. Method I shows the greatest departure from the group with the largest value of Ψ_{max} (caption, fig. 4a) and in the shape of the isotherms above the heat source (fig. 5a). The physical reason for this is that method I transfers more heat by convection from the heat source than the other methods (and thus has the greatest rate of fluid circulation, Ψ_{max}). This is discussed further at the end of section 4.3.

Figures 4 and 5 illustrate flows in which both convective and diffusive transport exists. The contribution of convection to the temperature fields in figure 5 can be appraised by referring to figure 6. This figure illustrates the steady state temperature field for static conduction (no fluid motion) and was obtained with method V by setting $Gr = U = V = \Omega = \Psi = 0$ in the basic differential equations (1)–(4). The conduction fields calculated in this way for numerical methods I–V are identical (to at least four significant figures) because all methods approximate the heat diffusion (or conduction) terms to the same spatial accuracy.

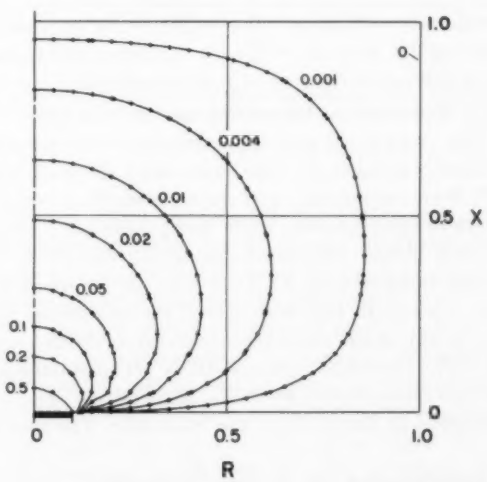


FIGURE 6. *Static conduction temperature field at steady state, all methods.*
Temperature (Θ) is the curve parameter; the heat source (thick line on floor) is at $\Theta = 1$.

Table 3 lists the computer time (in seconds) required to calculate the flows illustrated in figures 4 and 5. The particular computer used had execution times of 1.75 and 2.625 μs for nine digit floating point addition and multiplication, respectively. The only significant observation that can be made about this table is that method III required the least computer time, and method I the most.

Differences between methods could be reduced by altering the safety factor used for the time step $\Delta\tau$ or by more efficient programming.

TABLE 3. *Computer time required for test case with* $Gr = 1 \times 10^5$ *and a* 21×21 *mesh*

Method	Computer time
	sec
I	335
II	270
III	230
IV	309
V	285

4.3. Heat Transfer Rates

The heat transfer rates calculated by the five numerical methods are conveniently discussed in terms of the rate of heat addition ($\dot{Q}_{in}$) or heat removal ($\dot{Q}_{out}$) from the enclosure. These are total quantities, obtained by integration over the areas of the heat source and cold walls, respectively. The integration is performed by computing the heat transfer by convection and conduction from wall mesh points to adjacent mesh points, as discussed in the paragraph at the end of section 3.2a. A dimensionless heating rate can be defined as

$$\Phi = \frac{\dot{Q}}{\lambda a \Delta T} \tag{21}$$

where λ is the thermal conductivity of air, a is the height of the enclosure, and ΔT is the imposed temperature difference driving the flow, $\Delta T = T_h - T_0$. Equation (21) defines a dimensionless heat addition (Φ_{in}) or heat removal (Φ_{out}) as the rate of heat transfer is $\dot{Q}_{in}$ or $\dot{Q}_{out}$, respectively.

The heat transfer rates discussed in this section pertain to a ramp temperature change at the edge of the heat source. This ramp is the grid approximation of the boundary conditions (6) (a step change) and appears as a linear variation of temperature along the floor from $\Theta = 1$ at $R = R_b - \Delta R/2$ to $\Theta = 0$ at $R = R_b + \Delta R/2$. Both numerically and experimentally, a step change in temperature is difficult to achieve. The rate of heat transfer, while being finite for a ramp, would be infinite for a step. Twenty-one radial mesh points were used throughout this study and an extension of this study [18]. The corresponding ramp (see fig. 4 of [18]) is a close approximation of the floor temperature profile in a physical experiment [17] with which the numerical flows could be compared.

Curves illustrating the heat transfer rates for numerical methods I and II as a function of time are presented in figure 7; similar results for methods III–V are presented in figure 8. Figures 7 and 8 illustrate nonconservation and conservation methods, respectively. The ordinate is the dimensionless heat transfer rate into (Φ_{in}) or out of (Φ_{out}) the enclosure. The abscissa is the dimensionless time τ.

The calculated heat transfer rates due to static conduction ($Gr = 0$) are virtually identical for all five numerical methods. These conduction curves are shown in both figures 7 and 8 as a basis for comparison. Note that the heat removed (dashed line) approaches the heat added (solid line) with increasing time. Thus, the finite-difference approximation of the diffusion terms for all methods conserves energy within the grid system.

Methods I and II, figure 7, employ difference approximations of the convection terms which do not conserve energy. For method I, the rate of heat removal exceeds the rate of heat addition at steady state, indicating a net production of heat within the enclosure. For method II, the heat addition exceeds the heat removal, indicating a net absorption of heat within the enclosure. Clearly,

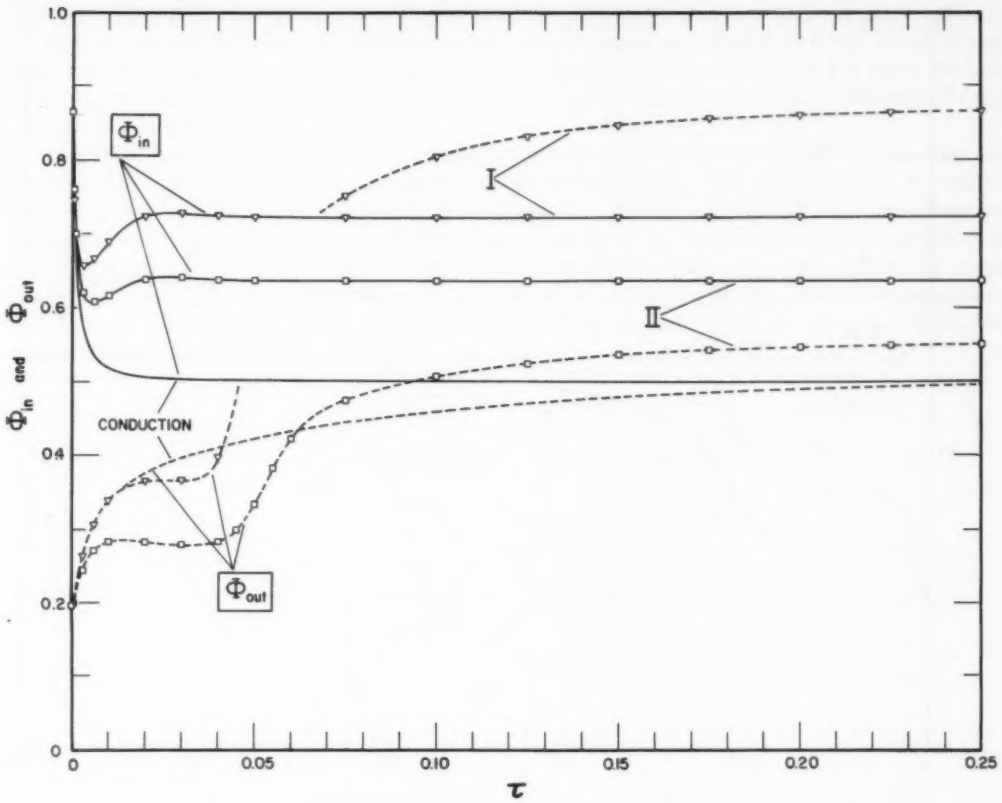


FIGURE 7. *Dimensionless heat flow into* (Φ_{in}) *and out of* (Φ_{out}) *the enclosure as a function of time* (τ) *for methods I and II* ($Gr = 1 \times 10^5$) *and static conduction (all methods).*

methods I and II violate the laws of thermodynamics. These differences in the rates of heat addition and heat removal are due to residues which remain when the difference equations for temperature are summed over all interior grid points, and the convection terms do not cancel in pairs. Methods I and II thus do not retain the energy conservation expressed by the basic differential equations.[8]

Methods III through V, figure 8, employ difference approximations of the convection terms which conserve energy. Clearly, for each method the rate of heat addition equals the rate of heat removal at steady state. Methods III and IV employ derivative approximations with similar spatial accuracy. As a consequence, the heat transfer curves for these methods are virtually identical.

Two general observations about figures 7 and 8 can be made. First, the shape of the heat addition curves is qualitatively similar for all methods, as is the shape of the heat removal curves. All curves for heat addition (solid lines) follow the conduction curve initially, then break away and achieve a steady value by $\tau = 0.05$. The heat removal curves (dashed lines) also follow the conduction curve initially. The heat removal rate then falls below the conduction value as it becomes more difficult for heat to be conducted radially from the heat source to the nearby cold floor counter to the incoming flow. At about $\tau = 0.04$, the rate of heat removal begins to increase sharply as the

[8] The authors of methods I [2] and II [20] did not calculate the heat transport from node to node within the mesh system. Instead, four point difference formulas were employed to approximate the temperature gradient at the walls, and thus the rates of heat transfer. For unsymmetric heating, this approach does not lead to equality between the steady state rates of heat removal and heat addition.

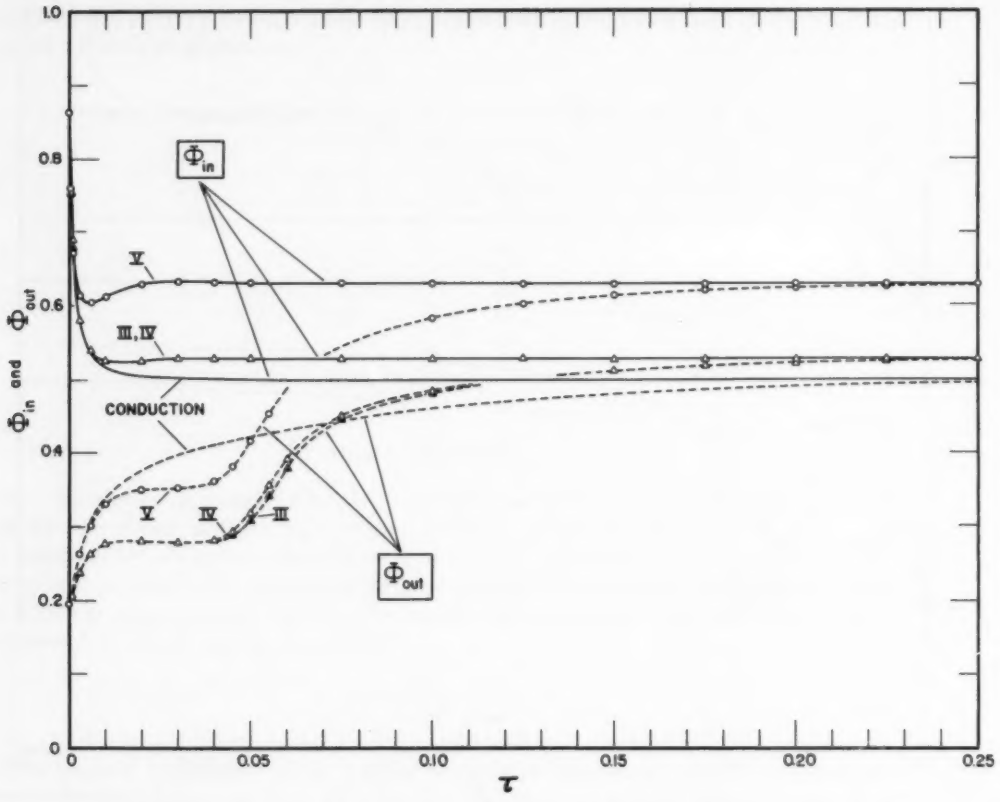


FIGURE 8. *Dimensionless heat flow into* (Φ_{in}) *and out of* (Φ_{out}) *the enclosure as a function of time* (τ) *for methods III, IV and V* ($Gr = 1 \times 10^5$) *and static conduction (all methods).*

heated fluid reaches the ceiling above the hot spot. As the fluid spreads radially outward along the ceiling, the rate of heat removal continues to rise. The lower curve then gradually approaches the upper (in fig. 8).

The second general observation that can be made about figures 7 and 8 concerns the magnitude of the steady-state heat additions. Clearly, the difference between the values for the five methods at $Gr = 1 \times 10^5$ and the value for conduction is a measure of the convective heat transfer from the heat source. This convective contribution decreases for the five methods in the sequence: method I > II > V > III = IV. The relative magnitudes of these contributions can be estimated by applying the difference approximation of the convection terms $U\partial\Theta/\partial X$ or $\partial(U\Theta)/\partial X$ (table 2) to the grid point $(i=1, j=0)$. A portion of this difference approximation is associated with the convective transport of energy from grid point $(i=0, j=0)$ to $(i=1, j=0)$. A careful study reveals that the convective heat transport for the five methods is proportional to the following quantities: method I $\propto U_{1,0}$; II $\propto U_{1,0}(1+\Theta_{1,0})/2$; V $\propto U_{1,0}/2$; and III and IV $\propto U_{1,0}\Theta_{1,0}/2$. (The boundary conditions $U_{0,0}=0$ and $\Theta_{0,0}=1$ were employed.) The temperature $\Theta_{1,0}$ lies between 0 and 1; thus, the aforementioned sequence of methods is in order of decreasing convection. This agrees with the results in figures 7 and 8.

4.4. Extension to Higher Gr, Convergence

The calculation of flows at high Grashof numbers is considered in this section. Results at $Gr = 1 \times 10^5$ in sections 4.1 and 4.2 revealed a qualitative similarity in the calculated flows for the

five numerical methods. Extension to higher Gr, however, means that the methods must continue to produce physically meaningful solutions, and this is closely associated with the practical stability, truncation errors, and conservation properties discussed in sections 3 and 4.3. For conservation reasons alone, we discard methods I and II from further consideration, and consider only the conservation methods III–V.

Of the latter methods, III and IV impose a spatial mesh size restriction to achieve stability while V does not. Violation of this restriction leads to oscillations in the fields of Θ, Ψ, U, and V. The effect of violating the spatial mesh size restriction is illustrated in figure 9 for the steady-state centerline temperature at $Gr = 1 \times 10^6$. The ordinate is the axial coordinate X and the abscissa is the temperature Θ. The curves on the right pertain to method III with the number of vertical mesh points, M, equal to 21, 41, and 61. (Results for methods III and IV were identical.) For comparison, results from method V are shown on the left with $M = 21$, 51, and 101. In all cases, the number of radial mesh points, N, is 21.

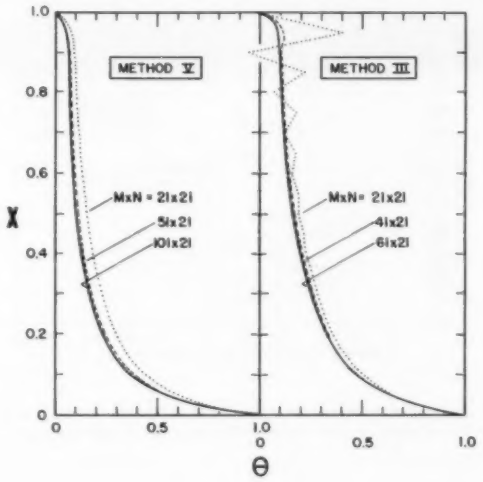


FIGURE 9. *Steady state distribution of temperature* (Θ) *with height* (X) *along the centerline; methods III and V*, Gr $= 1 \times 10^6$. Curve parameter is the number of vertical mesh points (M).

From figure 9, it is immediately apparent that method III leads to an oscillating temperature near the ceiling ($X = 1$) when $M = 21$. Repeating the calculation with $M = 41$ reduces the oscillation, but a local maximum persists at $X = 0.95$. Further refinement to $M = 61$ eliminates this hot spot. With 21 and 41 vertical mesh points, occasional negative temperatures were calculated during the transient. The steady state results for $M = 21$ reveal a region of negative temperature at $X = 0.8$. With $M = 11$, register overflows occurred early in the transient, and a stable integration was not possible.

With method V, no such oscillations or negative temperatures were calculated. Increasing the number of vertical mesh points from 21 to 101 reveals that the temperature distribution appears to asymptotically approach a limiting curve. Note that the asymptotic temperature distributions for methods III and V are not identical; this is attributed to differences in the truncation errors.

Oscillations such as those in figure 9 are physically unrealistic, and appear only when the mesh size restrictions (inequalities (15a)) are not satisfied. Such oscillations have not been observed with method V. The inequalities (15a) are satisfied for the test case at $Gr = 1 \times 10^5$ with a 21×21 mesh. As the Grashof number is increased, it has been found for the present problem [18] that $U_{max} \propto Gr^{1/2}$. Clearly, through (15a), this quickly leads to a very fine grid and prohibitively large computer storage requirements.

321-453 OL-69—4

The freedom of method V from a spatial mesh size restriction (and from oscillations) suggests its application to high Gr flows. Due to truncation errors which increase with Gr, such results must be interpreted carefully. The solutions should be tested for convergence or compared with physical experiment; two such tests for method V are discussed in the subsequent paragraphs.

A type of convergence check is illustrated in figure 10. The ordinate is the steady state heat flow into the enclosure, Φ_{in}, and the abscissa is the number of vertical mesh points, M. The curve parameter is the Grashof number. Solid lines pertain to method V, a single dashed line is shown for method III. For reasons noted at the beginning of section 4.3, the number of radial mesh points was held fixed at $N=21$ (except for the one solid data point for which $N=31$). The dotted line is an analytical result [8] for conduction of heat from a heated disk into a semi-infinite medium. The disk has a temperature profile quite close to the ramp profile of the grid approximation. The numerical results for static conduction are reasonably close to the analytical conduction value. For convective flows at $Gr=1\times10^5$, 1×10^6, and 4×10^7, the numerical results tend toward an asymptote with increasing M.

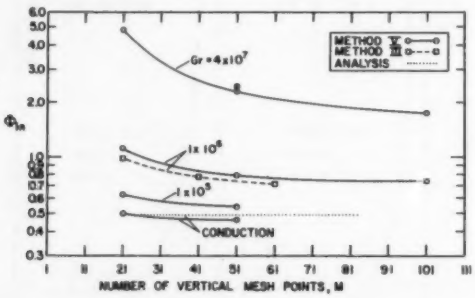


FIGURE 10. *Steady state dimensionless heat flow* (Φ_{in}) *versus number of vertical mesh points* (M).
Number of radial mesh points (N) held fixed at $N=21$ (except for one solid data point for which $N=31$).

Additional support for the use of method V is provided by comparison of a numerically calculated streamline field with physical experiment. An example of such a comparison is presented in figure 11 at $Gr=4\times10^6$. The numerical results are from an extension of this study [18], which employed a 51×21 grid ($M\times N$) and covered a range of Gr from 4×10^4 to 4×10^{10}. The experimental results are from reference [17] which covered a range of Gr from 8×10^5 to 1×10^{10}. For Gr up to about 1×10^9, the comparison between numerical and physical experiment is very good. Above this value, the physical experiment indicates turbulence. The numerical experiment was laminar (presumably due to false viscosity and the assumed two-dimensional motion) up to $Gr=4\times10^{10}$. At this value, the numerical flow developed a periodic vortex shedding, suggestive of the onset of laminar instability. This vortex shedding was adequately resolved with the grid spacing used.

A significant consideration in the application of any finite-difference procedure to a flow problem is the amount of computer time required to reach a steady state. Method V was applied in the numerical study of reference [18], and the computing time is shown in figure 12. Clearly, as Gr increases, the amount of computing time also increases (as $Gr^{1/4}$ at large Gr). Approximately one-half second was required for each time step, but for large Gr the size of the time step decreases rapidly and many more steps are required.

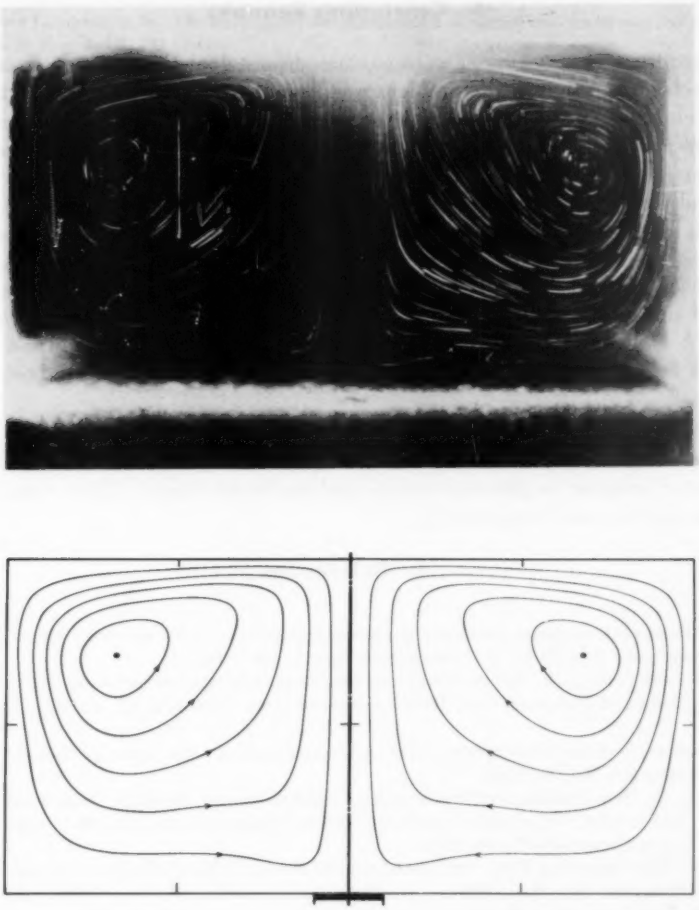


FIGURE 11. *Comparison of numerical results of method V (from reference [18]) with physical experiment (from reference [17]) at* $Gr = 4 \times 10^6$.

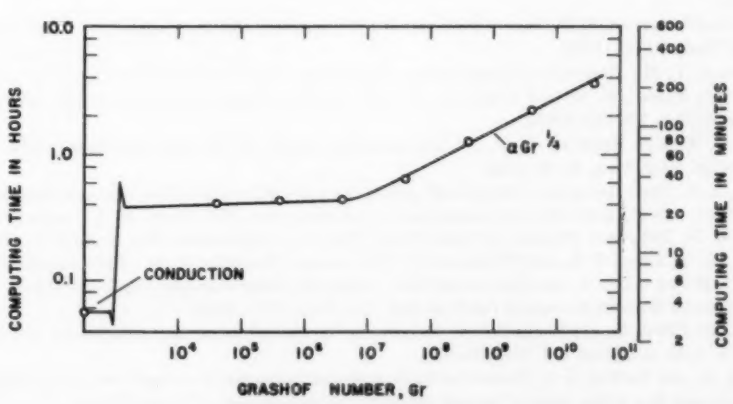


FIGURE 12. *Computing time required for method V at several Grashof numbers* (Gr), *numerical study of reference [18].*
Calculations carried from quiescent initial conditions to steady state with a $M \times N = 51 \times 21$ grid.

5. Concluding Remarks

Five finite difference procedures were compared for calculating transient natural convection flows in an enclosure. Although the problem was formulated in axisymmetric cylindrical coordinates, the conclusions should be applicable to any two-dimensional coordinate system.

To achieve stability of the time and space integrations, restrictions are imposed on the size of the time step for all methods (i.e., I–V) and on the size of the mesh spacing for methods II–IV.

Energy and vorticity are not conserved within the grid system with methods I and II, but are conserved with methods III–V. For this reason, the latter methods are to be preferred.

Of the conservation methods, method III appears to require less computer time than methods IV and V (table 3). Method III is therefore preferable when its associated mesh size restriction can be satisfied. If this restriction cannot be satisfied (which is usually the case) oscillations in the flow develop. A method free of a mesh size restriction (and oscillations) is then recommended, method V. The flows calculated with method V must be interpreted carefully, however, due to the truncation errors.

The author is indebted to L. Orloff, Factory Mutual Research Associate at the NBS, for developing a computer program to plot streamlines and isotherms directly from computed numerical fields of stream function and temperature.

6. References

[1] Arakawa, A., Computational design for long-term numerical integration of the equations of fluid motion: two-dimensional incompressible flow. Part 1, J. Computational Phys. **1**, 119 (1966).

[2] Barakat, H. Z., and Clark, J. A., Analytical and experimental study of the transient laminar natural convection flows in partially filled liquid containers, Proc. Third International Heat Transfer Conf. (Chicago, Illinois) Vol. **II**, 152 (1966).

[3] Bryan, K., A scheme for numerical integration of the equations of motion on an irregular grid free of nonlinear instability, Monthly Weather Rev. **94**, 39 (1966).

[4] Chandrasekhar, S., Hydrodynamic and Hydromagnetic Stability (Oxford University Press, London, England, 1961).

[5] DuFort, E. C., and Frankel, S. P., Stability conditions in the numerical treatment of parabolic equations, Math. Tables and Other Aids to Computation **7**, 135 (1953).

[6] Fromm, J., The Time Dependent Flow of an Incompressible Viscous Fluid, Methods in Computational Physics **3**, 345 (Academic Press, New York, N.Y., 1964).

[7] Hellums, J. D., and Churchill, S. W., Computation of natural convection by finite difference methods, International Developments in Heat Transfer—Proc. of the 1961–62 International Heat Transfer Conf. (Boulder, Colorado and London, England), 985 (1963).

[8] Karush, W., and Young, G., Temperature rise in a heat-producing solid behind a surface defect, J. Appl. Phys. **23**, 1191 (1952).

[9] Lax, P. D., and Richtmyer, R. D., Survey of the stability of linear finite difference equations, Communications on Pure and Appl. Math. **9**, 267 (1956).

[10] Milne-Thomson, L. M., Theoretical Hydrodynamics (Macmillan, New York, N.Y., fourth edition, 1960).

[11] O'Brien, G. G., Hyman, M. A., and Kaplan, S., A study of the numerical solution of partial differential equations, J. Math. and Phys. **29**, 223 (1950).

[12] Peaceman, D. W., and Rachford, H. H., Jr., The numerical solution of parabolic and elliptic differential equations, J. Soc. Indust. Appl. Math. **3**, 28 (1955).

[13] Richardson, L. F., The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam, Phil. Trans. R. S. London **210A**, 307 (1910).

[14] Richtmyer, R. D., Difference Methods for Initial Value Problems (Interscience, New York, N.Y., 1957).

[15] Runchal, A. K., Spalding, D. B., and Wolfshtein, M., The numerical solution of the elliptic equations for transport of vorticity, heat and matter in two-dimensional flows, Imperial College (London), Dept. Mech. Eng., SF/TN/2 (1967).

[16] Todd, J., Survey of Numerical Analysis (McGraw-Hill, New York, N.Y., 1962).

[17] Torrance, K. E., Orloff, L., and Rockett, J. A., Experiments on natural convection in enclosures with localized heating from below, to be published in J. Fluid Mech.

[18] Torrance, K. E., and Rockett, J. A., Numerical study of natural convection in an enclosure with localized heating from below—creeping flow to the onset of laminar instability, to be published in J. Fluid Mech.

[19] Wilkes, J. O., The finite difference computation of natural convection in an enclosed rectangular cavity (Ph. D. thesis, Univ. Michigan, Ann Arbor, 1963).

[20] Wilkes, J. O., and Churchill, S. W., The finite-difference computation of natural convection in a rectangular enclosure, Am. Inst. Chem. Eng. J. **12,** 161 (1966).
[21] Young, D., Iterative methods for solving partial difference equations of elliptic type, Trans. Am. Math. Soc. **76,** 92 (1954).

(Paper 72B4–279)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 72B, No. 4, October–December 1968

Generalized Inverses and Solutions of Linear Systems*

John Z. Hearon**

(September 28, 1968)

For an arbitrary complex matrix A we consider (1) the set of all matrices B such that $ABA=A$ and AB is Hermitian and (2) the set of all matrices B such that $ABA=A$ and BA is Hermitian. It is shown that if B is in (1) then $x=By$ is a least-squares solution of $Ax=y$ and that if B is in (2) then $x=By$ is the solution of minimum Euclidian norm of the consistent systems $Ax=y$. The connection is exposed between the properties of the generalized inverses in (a) and (b) and the fact that among all matrices X satisfying $AXA=A$, that with minimum Euclidian norm is the Moore-Penrose inverse of A.

Key Words: Generalized inverse; linear systems; matrix.

1. Introduction

In this paper we discuss some properties of certain classes of generalized inverses, give some new results and some enlightening connections between some known results.

It is shown how to construct (a) the set of all matrices B such that $ABA=A$ and AB is Hermitian and (b) the set of all matrices B such that $ABA=A$ and BA is Hermitian. The generalized inverses in (a) include the normalized generalized inverse of Rohde [15][1] and have been employed by Decell and Odell [3] in constructing the fixed point probability vector of regular or ergodic transition matrices. The generalized inverses in (b) include the Goldman-Zelen weak generalized inverse [4] and these authors have utilized a property of their inverse which we mention in the next paragraph.

It is well known [12] that given $Ax=y$, with A and y a given matrix and vector respectively, $x_0=A^+y$, where A^+ is the Moore-Penrose generalized inverse, is the least-squares solution of minimum Euclidian norm; in particular, the solution of minimum norm when the system is consistent. In what follows, we show that if B is in (b), then $x_0=By$ is the solution of minimum Euclidian norm when the system is consistent. It is further shown that if B is in (a), then $x=By$ is a least-squares solution of the system. We then show that the only matrix B such that $x_0=By$ is the least-squares solution of minimum norm is $B=A^+$. This is known [12]. The proof here shows the connection between this property and the weaker classes of generalized inverses mentioned above. All of the proofs assign a definitive role to a well-known property of perpendicular projections. The connection is shown between the properties of the generalized inverses in (a) and (b) and the fact that among all matrices B such that $ABA=A$, that with minimum Euclidian norm is A^+.

2. Definitions and Preliminaries

We consider matrices with complex entries. For any matrix M we denote by $\rho(M)$, $R(M)$, $N(M)$ and M^* the rank, range, nullspace and conjugate transpose respectively of M. By the norm, $\|M\|$, of any matrix M, we mean the Euclidian norm: $\|M\|^2=\text{trace}\,(M^*M)=\text{trace}\,(MM^*)$. For any column vector x, $\|x\|^2=x^*x$, and we observe that $\|M\|^2$ is the sum of the squared norms of the column vectors of M.

---

*An invited paper.

**Mathematical Research Branch, NIAMD, National Institutes of Health, Bethesda, Md. 20014.

[1] Figures in brackets indicate the literature references at the end of this paper.

We use the terminology for generalized inverses previously employed [7, 8, 9]. For a given matrix A, $C_1(A)$ is the set of all matrices B such that $ABA=A$; $C_2(A)$ is the set of all matrices B such that $B\epsilon C_1(A)$ and $BAB=B$; $C_3(A)$ is the set of all matrices $B\epsilon C_2(A)$ such that AB is Hermitian; $C_{3'}(A)$ is the set of all matrices $B\epsilon C_2(A)$ such that BA is Hermitian; $C_4(A)$ is the set of all matrices B such that $B\epsilon C_3(A)$ and $B\epsilon C_{3'}(A)$. We call a matrix $B\epsilon C_i(A)$ a C_i-inverse of A, $i=1, 2, 3, 3', 4$. We note that $C_4(A)$, the intersection of $C_3(A)$ and $C_{3'}(A)$, is a single matrix, the unique Moore-Penrose generalized inverse [11]. We observe that if $B\epsilon C_i(A)$ then AB and BA are projections and that $\rho(AB)=\rho(BA)=\rho(A)$ [15]. In particular, if i is 3, $3'$ or 4, then AB, BA or both, respectively, are perpendicular projections. We reserve the notation A^+ for the C_4-inverse of any matrix A.

We consider the system of equations

$$Ax=y, \tag{1}$$

where A and y are given. If $y\epsilon R(A)$, i.e., the system is consistent, then it is known [6] that there exists a solution $x_0\epsilon R(A^*)$, which is uniquely determined and is the solution of minimum norm: $\|x_0\|\leq\|x\|$, where x is any solution, with equality only if $x=x_0$. The proof is rather simpler than that [6] cited. Any solution x can be written $x=u+v$, with $u\epsilon R(A^*)$ and $v\epsilon N(A)$. Then clearly u is a solution. Let $x_1=u_1+v_1$ and $x_2=u_2+v_2$ be any two solutions with $u_i\epsilon R(A^*)$, $v_i\epsilon N(A)$, $i=1, 2$. Then $x_1-x_2=(u_1-u_2)+(v_1-v_2)$ is in $N(A)$, which implies that $(x_1-x_2)-(v_1-v_2)=u_1-u_2$ is in $N(A)$. But $(u_1-u_2)\epsilon R(A^*)$ and hence $u_1=u_2$. Thus any solution can be written $x=x_0+v$, with $x_0\epsilon R(A^*)$ uniquely determined and $v\epsilon N(A)$. Clearly $\|x\|^2=\|x_0\|^2+\|v\|^2\geq\|x_0\|^2$. In what follows, we therefore refer to *the* solution of minimum norm of a consistent system.

It is well known that if (1) is inconsistent, then any solution of

$$A^*Ax=A^*y \tag{2}$$

is a least-squares solution of (1). For, let x be any vector, x_1 satisfy (2), $d=Ax-y$ and $d_1=Ax_1-y$. Then, by (2), $d_1^*A=0$ and since $d_1-d=A(x_1-x)$ we have $d_1^*(d_1-d)=0$. Thus $d_1^*d_1=d_1^*d=d^*d_1$. From the Cauchy inequality, $(d^*d)(d_1^*d_1)\geq(d_1^*d)^2$, we then have $d^*d\geq d_1^*d_1$, with equality if and only if $d=d_1$. If equality is the case, then $d^*A=0$ and x is also a least-squares solution. Conversely, if x and x_1 are least-squares solutions then $d=d_1$, for then $(x-x_1)\epsilon N(A^*A)=N(A)$ which implies $Ax=Ax_1$. Thus every solution of (2) gives the same vector of deviations $d=Ax-y$, and, of necessity, the same minimum sum of squares. By what has been said in the last paragraph above, there exists a unique least-squares solution of minimum norm. Penrose [12] calls this least-squares solution the "best approximate solution" of (1), but he applies this concept and definition to more general systems than (1).

If (1) is consistent, any solution is a least-squares solution (zero sum of squares) and the solution of minimum norm is the least-squares solution of minimum norm. Thus the Penrose definition of "best approximate solution" is meaningful, and always exists for the system (1), without reference to the consistency of (1).

Lemma 1 gives a well-known property of perpendicular projections and Lemma 2 is an easy consequence of Lemma 1. We include both for ready reference and clarity.

LEMMA 1. *Let* E *be a projection. Then* $\|x\|\geq\|Ex\|$, *for every* x, *if and only if* $E=E^*$. *In that case,* $\|x\|=\|Ex\|$ *if and only if* $x\epsilon R(E)$.

PROOF. The first statement is a known theorem [5, p. 147]. To see the second statement, let $x=u+v$ with $u\epsilon R(E)$ and $v\epsilon N(E)$. Then $\|Ex\|^2=x^*Ex=u^*u$. But $\|x\|^2=x^*x=u^*u$ if and only if $x\epsilon R(E)$.

LEMMA 2. *Let* X *be any matrix and* E *and* F *be Hermitian projections. If* $X_0=EXF$, *we have* $\|X_0\|\leq\|X\|$, *with equality if and only if* $X=X_0$.

PROOF. Define $Y=EX$. Then by Lemma 1, $\|y_i\|\leq\|x_i\|$ where y_i and x_i are the ith column of Y and X respectively. Thus we have $\|Y\|\leq\|X\|$ with equality if and only if $Y=X$. Now $X_0^*=FY^*$, and by the same argument $\|X_0\|\leq\|Y\|$, with equality if and only if $X_0=Y$. Hence, $\|X_0\|\leq\|Y\|\leq\|X\|$, with equality only if $X_0=X$.

3. Generalized Inverses and Solutions of Linear Systems

In what follows we will be interested in and make use of matrices $B \epsilon C_1(A)$ such that either AB or BA is Hermitian. The next lemma gives conditions which such C_1-inverses must satisfy and in fact a method of constructiong them.

LEMMA 3. *Let* A *be an arbitrary, fixed matrix. Then every solution,* B, *of*

$$A^*AB = A^* \tag{3}$$

is in $C_1(A)$ *and is such that* AB *is a uniquely determined Hermitian projection. Conversely, every* $B \epsilon C_1(A)$ *such that* AB *is Hermitian satisfies* (3). *Similarly every solution,* B, *of*

$$AA^*B^* = A \tag{4}$$

is in $C_1(A)$ *and such that* BA *is Hermitian. Conversely every* $B \epsilon C_1(A)$ *such that* BA *is Hermitian satisfies* (4).

PROOF. Define $E = AB$. Then (3) reads $A^*E = A^*$ and left multiplication by B^* gives $E^*E = E$, which shows $E = E^* = E^2$. But then (3) can be written $A^*B^*A^* = A^*$ which implies that $B \epsilon C_1(A)$. Since $E = E^*$ and it is known that $\rho(A) = \rho(E)$, we have $R(E) = R(A)$ and E is completely determined by A. Conversely, if $ABA = A$ and AB is Hermitian we have $B^*A^*A = A$ which gives (3). The statements regarding the solutions of (4) are proved in the same way.

In the proof of his Theorem 1, Penrose [11] includes the second part of Lemma 3. Bose [2] has previously shown, for the real case, that if B is any solution of (3) then AB is a unique Hermitian projection. The proof is quite different (cf. Rohde [14]) and neither Bose nor Rohde drew the conclusion that $B \epsilon C_1(A)$. As before ([9]), we call a matrix B a *strict* C_1-inverse of A if $B \epsilon C_1(A)$ and $B \notin C_2(A)$, or equivantly ([7, 15]), if $B \epsilon C_1(A)$ and $\rho(B) > \rho(A)$. There exist strict C_1-inverses of A which satisfy (3) or (4) or both. For, if we write

$$A = Q \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} R \tag{5}$$

with Q and R unitary, where D is square, nonsingular and $\rho(A) = \rho(D)$, then [9] any $B \epsilon C_1(A)$ must have the form

$$B = R^* \begin{bmatrix} D^{-1} & U \\ V & W \end{bmatrix} Q^* \tag{6}$$

where the entries of U, V, and W are arbitrary. It is readily verified that AB is Hermitian if and only if $U = 0$; that BA is Hermitian if and only if $V = 0$; and that when either or both of these conditions are satisfied, $\rho(B) > \rho(A)$ if $W \neq 0$. Thus there exist solutions of (3) which are not in $C_3(A)$, solutions of (4) not in $C_{3'}(A)$ and solutions of both (3) and (4) other than the C_4-inverse of A.

The generalized inverses required by the theorem of Decell and Odell [3] are precisely those matrices satisfying (3), i.e., those matrices of the form (6) with $U = 0$.

Penrose [11] has shown that $KXM = P$ is consistent if and only if $P = KLPNM$, where L is any matrix in $C_1(K)$ and N is any matrix in $C_1(M)$. In that case, $X_0 = K^+PM^+$ is the solution of minimum norm [12]. The following theorem shows that, in the consistent case, the solution of minimum norm can be obtained by the use of generalized inverses in a weaker class than the C_4-inverse.

THEOREM 1. *Let* K, M, *and* P *be given matrices such that* KXM = P *is consistent. Then the solution of minimum norm is* $X_0 = LPN$, *where* L *is any matrix in* $C_1(K)$ *such that* LK *is Hermitian and* N *is any matrix in* $C_1(M)$ *such that* MN *is Hermitian.*

PROOF. Let X be any solution. Then $X_0 = LPN = LKXMN$ is also a solution. Since LK and MN are Hermitian projections, we have by Lemma 2, that $\|X_0\| \leq \|X\|$, with equality if and only if $X = X_0$.

Ben-Israel and Wersan [1] have shown that the solution of (3) with minimum norm is the C_4-inverse of A. We show how this follows from Theorem 1.

COROLLARY 1. *If* A *is a given matrix, the solution of* $A^*AX = A^*$ *of minimum norm is the* C_4-*inverse of* A.

PROOF. Let $K \in C_1(A^*A)$ and KA^*A be Hermitian. Then, by Theorem 1, $X_0 = KA^*$ is the solution of minimum norm. It is known [9] that $X_0 \in C_3(A)$. Since $X_0A = KA^*A$ is Hermitian, we have $X_0 \in C_4(A)$.

In view of Lemma 3, Corollary 1 shows that among all matrices $B \in C_1(A)$ such that AB is Hermitian, the C_4-inverse of A has minimum norm. In fact among *all* matrices in $C_1(A)$, the C_4-inverse of A has minimum norm. A proof of this has been indicated by Kalman [10]. We give this result below as a corollary of Theorem 1.

COROLLARY 2. *If* A *is a given matrix, the solution of* AXA = A *of minimum norm is the* C_4-*inverse of* A.

PROOF. Clearly $AXA = A$ is consistent. By Theorem 1, the solution of minimum norm is $X_0 = B_1AB_2$ where B_1 and B_2 are in $C_1(A)$ and are such that B_1A and AB_2 are Hermitian. By a known theorem [7], $X_0 \in C_2(A)$, and plainly $AX_0 = AB_2$ and $X_0A = B_1A$ are Hermitian. Hence $X_0 \in C_4(A)$.

REMARK. One construction of the C_4-inverse of A given by Penrose [11] is in fact $B = B_1AB_2$, where B_1 and B_2 are any solutions of (4) and (3), respectively. This is a special instance of the theorem [7] that $B = B_1AB_2$ is in $C_2(A)$ when B_1 and B_2 are *arbitrary* elements of $C_1(A)$. In particular, if AB_2 is Hermitian, then $B \in C_3(A)$ and if B_1A is Hermitian, then $B \in C_{3'}(A)$ ([cf. 9]). We could also prove Corollary 2 by observing that among all matrices of the form (6),

$$B = R^* \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q^*$$

has minimum norm and that, according to the discussion after (6), this matrix is A^+.

The next theorem gives a subset of $C_1(A)$ such that elements of this subset select the solution of minimum norm when (1) is consistent. It also gives a subset of $C_1(A)$ which selects a least-squares solution of (1) and shows that the C_4-inverse of A is the only matrix which enjoys both properties.

THEOREM 2. *Consider the system*

$$\mathrm{Ax} = \mathrm{y} \tag{s}$$

where A *and* y *are a given matrix and vector respectively.*

(i) Let $B \in C_1(A)$ and BA be Hermitian. If (s) is consistent, then $x_0 = By$ is the solution of minimum norm. Conversely, if $x_0 = By$ is the solution of minimum norm for every y such that (s) is consistent, then $B \in C_1(A)$ and BA is Hermitian.

(ii) Let $B \in C_1(A)$ and AB be Hermitian. Then $x = By$ is a least-squares solution. Conversely, if $x = By$ is a least-squares solution for every y, then $B \in C_1(A)$ and AB is Hermitian.

(iii) Let B be the C_4-inverse of A. Then $x_0 = By$ is the least-squares solution of minimum norm. Conversely, if $x_0 = By$ is the least-squares solution of minimum norm for every y, then B is the C_4-inverse of A.

PROOF. (i) It follows from Theorem 1 that $x_0 = By$ is the solution of minimum norm if $B \in C_1(A)$ and BA is Hermitian. Conversely, suppose $x_0 = By$ to be the solution of minimum norm for each $y \in R(A)$. Then $Ax_0 = ABy = y$, for all $y \in R(A)$ implies $ABAz = Az$, for arbitrary z, and so $B \in C_1(A)$. Thus far we have only used the fact that By is a solution. For any z, let $y = Az$, then we have $By = BAz$ and $\|By\| = \|BAz\|$. If By is the solution of minimum norm, $\|By\| = \|BAz\| \leq \|z\|$, and that BA is Hermitian follows from Lemma 1.

(ii) The least-squares solutions of (s) are the solutions of $A^*Ax = A^*y$. We show that $x = By$ with B as in (ii) satisfies this equation. We have $A^*Ax = A^*ABy = A^*B^*A^*y = A^*y$, since $AB = B^*A^*$ and $ABA = A$. Conversely, if $A^*ABy = A^*y$ for every y, we have $A^*AB = A^*$ and, from Lemma 3, $B \epsilon C_1(A)$ and AB is Hermitian.

(iii) First, let $B = A^+$. Then by (ii), x_0 is a least-squares solution. Now let x be any least-squares solution. Then $A^*Ax = A^*y$ from which $BB^*A^*Ax = BB^*A^*y$. But $B^*A^* = AB$ and we have $BAx = By = x_0$. Since BA is Hermitian, $\|x_0\| \leq \|x\|$ follows from Lemma 1. Conversely, suppose that for every y, $x_0 = By$ is the minimum norm solution of $A^*Ax = A^*y$. By Theorem 1 we then have $x_0 = By = KA^*y$ for all y, where $K \epsilon C_1(A^*A)$ and KA^*A is Hermitian. Thus $B = KA^*$, which as in the proof of Corollary 1 is the C_4-inverse of A.

REMARK. In the proof of (i) we have essentially proved that if $B \epsilon C_1(A)$, then $x = By$ is a solution of (s) whenever $y \epsilon R(A)$ and that, conversely, if $x = By$ is a solution of (s) for every $y \epsilon R(A)$ then $B \epsilon C_1(A)$. This was proved by Bose [2] (cf. [14]) and in fact Rao [13] takes this as a definition of $C_1(A)$.

The content of (iii) of Theorem 2 is known. For, Penrose [12] shows that $X_0 = A^+C$ is the least-squares solution of minimum norm of $AX = C$. By observing that A^+ is the least-squares solution of minimum norm of $AX = I$ he shows that A^+ is the only matrix with the required properties. We prove (iii) here to show how the result stems directly from the Hermitian character of A^+A and AA^+ and to show the relation of (iii) to (i) and (ii).

It is clear that any $B \epsilon C_{3'}(A)$ meets the conditions of (i) of Theorem 2. It is precisely that property of the $C_{3'}$-inverse which was exploited by Goldman and Zelen [4]. Their proof however is essentially different in character.[2] With Theorem 2 in hand, we can give a rather brief version of their proof of the Gauss theorem (Theorem 1 of [4]) utilizing a weaker class of generalized inverse than the $C_{3'}$-inverse. The term *estimable* is used here as it is used in [4].

THEOREM 3. Let y be a vector of random variables such that the variance-covariance matrix is $\sigma^2 I$ and the expected value of y is $A\zeta$. The minimum variance unbiased linear estimate $\hat{\theta}$ of any estimable function $\theta = q^*\zeta$ is given by $\hat{\theta} = q^*By = q^*x$, where B is any matrix in $C_1(A)$ such that AB is Hermitian and x is any solution of (2).

PROOF. A linear estimate $\hat{\theta} = p^*y$ of θ is unbiased if and only if the expected value $E(\hat{\theta})$ of $\hat{\theta}$ is $E(\hat{\theta}) = p^*A\zeta = \theta$ for all ζ, i.e., $A^*p = q$. Thus θ is estimable if and only if $q \epsilon R(A^*)$, and since the variance of $\hat{\theta}$ is $(p^*p)\sigma^2$, $\hat{\theta}$ gives the minimum variance unbiased estimate if and only if p is the minimum norm solution of the consistent system $A^*p = q$. According to (i) of Theorem 2, this is equivalent to $p = B^*q$ where $B^* \epsilon C_1(A^*)$ and B^*A^* is Hermitian, i.e., $B \epsilon C_1(A)$ and AB is Hermitian. Thus $\hat{\theta} = p^*y = q^*By$. But for any $q \epsilon R(A^*)$ we further have $q^*By = q^*x$, where x is any solution of (2). For, by (ii) of Theorem 2, By is a solution of (2) and any solution of (2) can be written $x = By + v$ with $v \epsilon N(A)$.

It is of interest to note that the general solution of (2) can be written as $x = KA^*y + (I - KA^*A)z$, where K is any C_1-inverse of A^*A and z is arbitrary.[3] This, in a different notation, is eq (3.3) of Goldman and Zelen [4], except that they require that $K \epsilon C_2(A^*A)$. As noted previously in this paper, and shown in [9], $KA^* \epsilon C_3(A)$. Thus Goldman and Zelen in fact employ both the C_3-inverse and the $C_{3'}$-inverse and these are included in the set of all matrices which meet the conditions of (ii) and (i) of Theorem 2, respectively.

3. References

[1] Ben-Israel, A., Wersan, S. J., An elimination method for computing the generalized inverse of an arbitrary complex matrix, J. Assn. for computing Machinery **10**, 532–537 (1963).

[2] Bose, R. C., Lecture Notes on Analysis of Variance, University of North Carolina, Chapel Hill, 1959.

[3] Decell, H. P., Jr., and Odell, P. L., The fixed point probability vector of regular or ergodic transition matrices, J. Am. Stat. Assoc. **62**, 600–602 (1967).

[2] Their argument is given immediately preceding their equation (3.6) and they make explicit use of the fact that $C_3(A)$ is contained in $C_{3'}(A)$.

[3] This is a straightforward application of Penrose's Corollary 1 [11] and we observe with Penrose that the only property of the C_4-inverse required for the proof of the corollary is $B \epsilon C_4(A)$ implies $B \epsilon C_1(A)$.

[4] Goldman, A. J., and Zelen, M., Weak generalized inverses and minimum variance linear unbiased estimation, J. Res. NBS **68B** (Math. and Math. Phys.), No. 4, 151–172 (1964).
[5] Halmos, P., Finite-dimensional Vector Spaces (D. Van Nostrand, New York, 1958).
[6] Hamburger, H. L., and Grimshaw, M. E., Linear Transformation in *n*-dimensional Vector Space (Cambridge University Press, 1951).
[7] Hearon, J. Z., Construction of *EPr* generalized inverses by inversion of nonsingular matrices, J. Res. NBS **71B** (Math. and Math. Phys.), Nos. 2 and 3, 57–60 (1967).
[8] Hearon, J. Z., A generalized matrix version of Rennie's inequality, J. Res. NBS **71B** (Math. and Math. Phys.), Nos. 2 and 3, 61–64 (1967).
[9] Hearon, J. Z., and Evans, J. W., On spaces and maps of generalized inverses, J. Res. NBS **72B** (Math. Sci.), No. 2, 103–107 (1968).
[10] Kalman, R. E., New Methods and Results in Linear Prediction and Filtering Theory, Technical Report 61–1, Research Institute for Advanced Studies, Baltimore, Md. (1960).
[11] Penrose, R., A generalized inverse for matrices, Proc. Cambridge Philos. Soc. **51,** 406–418 (1955).
[12] Penrose, R., On best approximate solutions of linear matrix equations, Proc. Cambridge Philos. Soc. **52,** 17–19 (1956).
[13] Rao, C. R., A note on a generalized inverse of a matrix with applications to problems in mathematical statistics, J. Royal Stat. Soc. (b) **24,** 152–158 (1962).
[14] Rohde, C. A., Contributions to the Theory, Computation and Application of Generalized Inverses, Ph.D. Thesis, North Carolina State University, Raleigh (1964).
[15] Rohde, C. A., Some results on generalized inverses, SIAM Rev. **8,** 201–205 (1966).

(Paper 72B4–280)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 72B, No. 4, October–December 1968

Publications of the National Bureau of Standards*

Selected Abstracts

Eisenhart, C., **Expression of uncertainties of final results,** *Science* **160,** *No. 3833, 1201–1204 (June 14, 1968).*
The terms used to describe the precision and accuracy of measurements are defined and discussed. Specific rules are recommended for expressing the uncertainty of a reported value for four cases distinguished according to the relative magnitudes of the imprecision and the likely systematic error of the measurement process employed. Examples of recommended forms of uncertainty statements are given.

Key Words: Accuracy; calibrations; expression of uncertainty; measurements; precision; standard error; systematic error; uncertainty.

Newman, M., **Pairs of matrices generating discrete free groups and free products,** *Mich. Math. J.* **15,** *155–160 (1968).*
It is shown that if

$$A=\begin{pmatrix}-a & b\\ -c & d\end{pmatrix}, \qquad B=\begin{pmatrix}-\alpha & -\beta\\ \gamma & \delta\end{pmatrix}$$

are elements of $SL(2, R)$ such that $a, b, c, d, \alpha, \beta, \gamma, \delta \geqslant$ and $d-a \geqslant 2$, $\delta-\alpha \geqslant 2$, than the group $\{A, B\}$ is a free discrete subgroup of $SL(2, R)$. Generalizations to free products are given.

Key Words: Free groups; discrete matrix groups; free products.

Osgood, C. F., **A method in diophantine approximation (II),** *Acta Arith.* **XIII,** *383–393 (1968).*
This paper applies the general method of approach developed in *A Method in Diophantine Approximation* to obtain further results about the diophantine approximation of values of certain functions. Examples of such functions are the inverse Laplace transforms of all functions of the form $z^{-\alpha}\exp\left(\sum_{i=1}^{l-1} r_i z^{-i}\right)$ where α and the r_i are each rational, $\alpha > l+1$, α is not an integer and $r_{l-1} \neq 0$. For these latter functions we conclude that at any nonzero rational point the zeroth through l-1st derivatives are linearly independent over the rationals.

Key Words: Integers; irrational numbers; diophantine approximation; differential equations; linear operators; and convolution product.

Other NBS Publications

J. Res. NBS 72A (Phys. and Chem.), No. 5 (Sept.–Oct. 1968).

Studies in bomb calorimetry. A new determination of the energy of combustion of benzoic acid in terms of electrical units. K. L. Churney and G. T. Armstrong.
Effect of low pressures on the room temperature transitions of polytetrafluoroethylene.. G. M. Martin and R. K. Eby.
Electrostatic potentials and their spatial derivatives about point defects in ionic crystals. H. S. Bennett.
Electric fields produced by a charge density in ionic crystals. H. S. Bennett.
Mass spectrometric study of the photoionization of some fluorocarbons and trifluoromethyl halides. C. J. Noutary.
Infrared matrix spectra of lithium fluoride. S. Abramowitz, N. Acquista, and I. W. Levin.
Electronic transition moment integrals for first ionization of CO and the A-X transition in CO^+. Some limitations on the use of the r-centroid approximation. P. H. Krupenie and W. Benesch.
The configurations $3d^84p$ in doubly ionized atoms of the iron group. C. Roth.
Electron impact excitation of hydrogen Lyman-α radiation. R. L. Long, Jr., D. M. Cox, and S. J. Smith.

J. Res. NBS 72A (Phys. and Chem.), No. 6 (Nov.–Dec. 1968).

Interaction energy surfaces for $Li(2^2S)$ and $Li(2^2P)$ with H_2. M. Krauss.
Energy levels and classified lines in the first spectrum of technetium (Tc I). W. R. Bozman, C. H. Corliss, and J. L. Tech.
The first spectrum of tungsten (W I). D. D. Laun and C. H. Corliss.
Theoretical interpretation of the even levels in the first spectrum of tungsten. Y. Shadmi and E. Caspi.
Stable carbon isotope ratio measurement with a gas density meter. S. P. Wasik and W. Tsang.
The single crystal spectrum of hexakis(imidazole)nickel(II) nitrate. C. W. Reimann.
Crystalline alpha and beta forms of 3-0-α-glucopyranosyl-D-arabinopyranose. H. S. Isbell, H. L. Frush, and J. D. Moyer.
Preparation and solubility of hydroxyapatite. E. C. Moreno, T. M. Gregory, and W. E. Brown.
High temperature dehydroxylation of apatitic phosphates. T. Negas and R. S. Roth.
Bates, R. G., Ed., Electrochemical Analysis Section: Summary of activities July 1967 to June 1968, Tech. Note 453 (July 1968), 55 cents.
Beam, A. E., Hilsenrath, J., PRECISE: A multiple precision version of omnitab, Tech. Note 446 (June 1968), 55 cents.
Brady, E. L., Editor, Status report National Standard Reference Data System, April 1968, Tech. Note 448 (June 1968), 70 cents.
Hall, L. A., A bibliography of thermophysical properties of methane from 0 to 300 °K, Tech. Note 367 (May 1968), 60 cents.
Horiuti, J., Miyahara, K., Hydrogenation of ethylene on metallic catalysts, NSRDS13 (June 1968), $1.00.
Kessler, K. G., Passaglia, E., Winogradoff, N. N., Research on laser standards and materials at the National Bureau of Standards, Tech. Note 449 (June 1968), 30 cents.
Koeser, R. L., Editor, Report of the 52nd National Conference on Weights and Measures 1967, Spec. Publ. 297 (May 1968), $1.25. (This series was formerly known as the Miscellaneous Publication Series.)
MacKay, D. R., Editor, Editorial format for Product Standards, PS0–67 (July 1968), 10 cents.
Mason, H. L., Editor, Calibration and test services of the National Bureau of Standards, Spec. Publ. 250, 1968 edition (May 1968), $1.75 (Supersedes Misc. Publ. 250, 1965 edition).
Moore, C. E., Merrill, P. W., Partial Grotrian diagrams of astrophysical interest, NSRDS–NBS23 (June 1968), 55 cents.
Mosburg, E. R., Jr., Lojko, M. S., Solution of the Abel integral transform for a cylindrical luminous region with optical distortions at its boundary, Tech. Note 368 (July 12, 1968), 25 cents.
Newman, M., Matrix representations of groups, AMS60 (July 1968), 60 cents.
Passaglia, E., Ed., Research on high temperature materials at the National Bureau of Standards, Tech. Note 447 (July 1968), 20 cents.
Roder, H. M., McCarty, R. D., Johnson, V. J., Saturated liquid densities of oxygen, nitrogen, argon, and parahydrogen, Tech. Note 361 (Jan. 31, 1968), 50 cents.

Rosenberg, S. J., Nickel and its alloys, Mono. 106 (May 1968), $1.25 (Supersedes Circular 592).

Schafft, H. A., Needham, S. G., A bibliography on methods for the measurement of inhomogeneities in semiconductors (1953–1967), Tech. Note 445 (May 1968), 35 cents.

Schwendeman, R. H., Tables for the rigid asymmetric rotor: Transformation coefficients from symmetric to asymmetric bases and expectation values of P_z^2, P_x^2, and P_y^2, NSRDS–NBS 12 (June 1968), 60 cents.

Tipson, R. S., Infrared spectroscopy of carbohydrates, a review of the literature, Mono. 110 (June 1968), 30 cents.

Acquista, N., Schoen, L. J., Lide, D. R., Jr., Infrared spectrum of the matrix-isolated OH radical, J. Chem. Phys. **48,** No. 4, 1534–1536 (Feb. 15, 1968).

Adams, J. W., Desch, R. F., Experimental confirmation of barretter substitution error, IEEE Trans. Microwave Theory Tech. MTT–16, No. 3, 201–202 (Mar. 1968).

Alley, C. O., Bender, P. L., Information obtainable from laser range measurements to a lunar corner reflector, Book, IAU-IUGG Symposium #32 on Continental Drift, Secular Motion of the Pole and the Rotation of the Earth, 32, 86–90 (Reidel Publ. Co., Dordrecht, The Netherlands, 1968).

Ambrose, J. R., Kruger, J., The stress-corrosion of Ti and Ti-8Al-1Mo-1V in methanol vapour, Corrosion Sci. **8,** 119–124 (1968).

Armstrong, R. W., NBS dynamic seat belt tester, Proc. 10th Stapp Car Crash Conf., 6571st Aeromedical Research Laboratory, Alamogordo, New Mex., Nov. 8–9, 1966, pp. 38–45 (Society of Automotive Engineers, New York, N.Y., 1966).

Astin, A. V., Standards of measurement, Sci. Am. **218,** No. 6, 50–62 (June 1968).

Astin, A. V., Voluntary standardization and the government (fifty years of cooperation), Mag. Stds. **39,** No. 6, 167–172 (June 1968).

Astin, A. V., Welcoming remarks (Proc. Fifth Transducer Workshop, National Bureau of Standards, Gaithersburg, Md., Oct. 3–4, 1967), Telemetry Working Group Document 112–68, 3–4 (Secretariat, Range Commanders Council, White Sands Missile Range, New Mexico, 1968).

Bates, R. G., Standardization of acidity measurements. Extension of the pH concept to mixed solvents and heavy water, Anal. Chem. 40, 28A–38A (May 1968).

Beckett, C. W., Status of properties research in the kilodegree range, Book, The Characterization of High-Temperature Vapors, Ed. J. L. Margrave, ch. 1, pp. 3–15 (John Wiley & Sons, Inc., New York, N.Y., 1967).

Bennett, J. A., To avoid fatigue failures, pay attention to details, Current Engr. Pract. **9,** 12–16 (Sept. 1966).

Berendt, R. D., Winzer, G. E., and Burroughs, C. B., A guide to airborne, impact, and structure borne noise control in multifamily dwellings, FT/TS–24 (U.S. Dept. Housing and Urban Development, Supt. Doc's., Government Printing Office, Wash., D.C., Jan. 1968, $2.50).

Billick, I. H., Dishon, M., Weiss, G. H., Yphantis, D. A., Numerical solutions of the Lamm equation. IV. Rotor slowing experiments, Biopolymers **5,** 1021–1028 (1967).

Billick, I. H., Schulz, M., Weiss, G. H., Quasi-equilibrium sedimentation experiments with rotor deceleration, J. Phys. Chem. **71,** No. 8, 2495–2502 (July 1967).

Bleicher, M. N., Knopp, M. I., Lattice points in a sphere, Acta Arith. X, 369–376 (1965).

Bloss, R. L., The measurement of strain (Proc. 22nd Annual ISA Conf., Sept. 11–14, 1967, Chicago, Ill.) IRA Preprint No. PI–1–PHYMMID–67 (1967).

Bollacasa, D., Goldman, D. T., Importance of upscattering in the calculation of neutron spectra, Proc. Intern. Atomic Energy Agency Symp. Neutron Thermalization and Reactor Spectra, Ann Arbor, Michigan, July 17–21, 1967, 1, 537–554 (IAEA, Vienna, Austria, Jan. 1968).

Bowman, H. A., Gallagher, W., and Schoonover, M. R., The development of a working density standard (Proc. 20th Annual ISA Conf., Oct. 4–7, 1965, Los Angeles, Calif.) ISA Preprint No. 14.8–4–65 (1965).

Bozman, W. R., The application of computers in printing scientific manuscripts, Book, New Techniques in Office Operations, Ch. 13, pp. 120–124 (The Business Press, Elmhurst, Ill, Mar. 1968).

Brady, G., McIntyre, D., Myers, M. E., Jr., and Wims, A. M., Critical scattering of perfluoroheptane-isooctane system (Proc. Conf. Small-angle X-ray Scattering, Syracuse Univ., Syracuse, N.Y., June 1965), Book, Small-Angle X-ray Scattering, Ed. H. Brumberger, pp. 425–448 (Gordon and Breach, Science Publ. Inc., New York, N.Y., 1967).

Branscomb, L. M., The role of atomic collision processes in astrophysics, Book, The Physics of Electronic and Atomic Collisions: Invited Papers from the Fifth International Conference, Leningrad, U.S.S.R., July 17–23, 1967, Ed. L. M. Branscomb, pp. 12–31 (Joint Institute for Laboratory Astrophysics, University of Colorado, Boulder, Colorado, 1968).

Branscomb, L. M., Twenty years of physics: atoms, molecules and electrons, Phys. Today **21,** No. 5, 36–39 (May 1968).

Braun, W., and Lenzi, M., Resonance fluorescence method for kinetics of atomic reactions. Reactions of atomic hydrogen with olefins, Disc. Far. Soc. **44,** 252–261 (1967).

Breen, D. R., Keller, R. A., Intramolecular energy transfer between triplet states of weakly interacting chromophores. I. Compounds in which the chromophores are separated by a series of methylene groups, J. Am. Chem. Soc. **90,** No. 8, 1935–1940 (Apr. 10, 1968).

Brenner, A., Metzger, W. H., Jr., Determination of the composition of complexes and their instability constants by calorimetery III. The complex in fused sodjum molybdate and molybdenum trioxide, J. Electrochem. Soc. **115,** No. 3, 258–261 (Mar. 1968).

Bridges, J. M., and Wiese, W. L., Comparison of the Stark widths of the Balmer lines H_β and H_α, Proc. Seventh Intern. Conf. Phenomena in Ionized Gases, Ed. B. Perovic and D. Tosic, III, No. 165, 165–167 (Gradevinska Knjiga Publ. Co., Beograd, Yugoslavia, 1966).

Brown, R. L., Brennen, W., Spin-relaxation effects on the EPR spectrum of gaseous nitrogen atoms, J. Chem. Phys. **47,** No. 11, 4697–4705 (Dec. 1, 1967).

Brown, W. E., A mechanism for growth of apotitic crystals (Proc. Intern. Symp. Composition, Properties and Fundamental Structure of Tooth Enamel, London, England, Apr. 1964), Book, Tooth Enamel, pp. 11–14 (John Wright & Sons, Ltd., Bristol, England, 1965).

Bullis, W. M., and Coleman, J. A., Characterization of germanium and silicon for nuclear radiation detectors, Book, Nucleonics in Aerospace, Ed. P. Polishuk, pp. 166–175 (Plenum Press Inc., New York, N.Y., 1968).

Burr, A. J., and Roberts, D. E., High polymeric materials, Digest of Literature on Dielectrics **28,** 231–260 (1964).

Butterfield, M. A., Care and preservation of the new media-equipment needs, Proc. A Pioneer Presentation of a National Symposium on the Impact of Automation on Documentation, Denver, Colo., Apr. 27–29, 1967, pp. 60–64 (1968).

Butterfield, M. A., Optical character recognition as an archival tool, Proc. A. Pioneer Presentation of a National Symposium on the Impact of Automation and Documentation, Denver, Colo., Apr. 27–29, 1967, pp. 101–118 (1968).

Caldwell, B. P., Kubelka-Munk coefficients from transmittance, J. Opt. Soc. Am. **58,** No. 6, 755–759 (June 1968).

Campbell, P. G., and Wright, J. R., Asphalt hardening by gaseious oxidants: the relationship betwcen softening point and chemical change as measured by infrared spectroscopy, J. Mater. **2,** No. 3, 581–596 (Sept. 1967).

Canfield, L. R., Johnson, R. G., Codling, K., and Madden, R. P., Comparison of an ionization chamber and a thermopile as absolute detectors in the extreme ultraviolet, Appl. Opt. **6,** No. 11, 1886–1888 (Nov. 1967).

Cassidy, E. C., time-resolved studies of spectra produced by electrically exploded wires, Naturwissenschaften 55, No. **3,** 125–128 (1968).

Catanzaro, E. J., Absolute isotopic abundance ratios of three common lead reference samples, Earth and Planetary Sci. Letters **3,** No. 4, 343–346 (1968).

Chang, S. S., Bestul, A. B., Heat capacities for atactic polystyrene of narrow molecular weight distribution to 360 °K, J. Polymer Sci. **6,** Part A–2, 849–860 (1968).

Childs, G. E., Hanley, H. J. M., Applicability of dilute gas transport property tables to real gases, Cryogenics **8,** No. 2, 94–97 (Apr. 1968).

Clark, J. E., and Harrison, C. W., Accelerated weathering of polymers: radiation, Appl. Polymer Symp. **4,** 97–110 (1967).

Cohen, M. I., Blunt, R. F., Optical properties of $SrTiO_3$ in the region

of the absorption edge, Phys. Rev. **168,** No. 3, 929–933 (Apr. 15, 1968).
Coleman, J. A., Germanium for gamma-ray detectors. A review of current problems, Book, Lithium-Drifted Germanium Detectors, STI/Publ./132, pp. 37–41 (International Atomic Energy Agency, Vienna, Austria, 1966).
Coleman, J. A., Love, D. P., Trainor, J. H., and Williams, D. J., Low-energy proton damage effects in silicon surface-barrier detectors, IEEE Trans. Nucl. Sci. **NS–15,** No. 1, 482–490 (Feb. 1968).
Cook, R. L., and Kirchoff, W. H., Further investigations on the microwave spectrum of NSF; evaluation of the molecular force field, centrifugal distortion constants, and the dipole moment, J. Chem. Phys. **47,** No. 11, 4521–4527 (Dec. 1, 1967).
Cooper, M. J., Comments on the scaling-law equation of state, Phys. Rev. **168,** No. 1, 183–184 (Apr. 5, 1968).
Cooper, M. J., Mountain, R. D., Interpretation of relaxation times in the "electric-field effect on critical opalescence," J. Chem. Phys. **48,** No. 3, 1064–1066 (Feb. 1, 1968).
Corliss, E. L. R., Burnett, E. D., Kobal, M. T., Bassin, M. A., The relative importance of frequency distortion and changes in time constants in the intelligibility of speech, IEEE Trans. Audio Electroacoustics **AU–16,** No. 1, 36–39 (Mar. 1968).
Corliss, E. L. R., Burnett, E. D., Stimson, H. F., Polyacusis, a hearing anomaly, J. Acoust. Soc. Am. **43,** No. 6, 1231–1236 (June 1968).
Cornell, R. G., Speckman, J. A., Estimation for a simple exponential model, Biometrics **23,** No. 4, 717–737 (Dec. 1967).
Cornog, J. R., Bryan, H. L., Jr., Manual and machine-assisted search methods used in examining patent applications in the transistor art, Proc. Natl. Electronics Conf., Chicago, Ill., Oct. 25–27, 1965, XXI, 672–677 (1965).
Costrell, L., Standard nuclear instrument modules (Adopted by AEC Committee on Nuclear Instrument Modules January 1968), AEC Report TID–20893 (Rev. 2), (Superintendent of Documents, Government Printing Office, Washington, D.C., 1968, 25 cents).
Covington, A. K., Paabo, M., Robinson, R. A., and Bates, R. G., Use of the glass electrode in deuterium oxide and the relation between the standardized pD (p_{a_D}) scale and the operational pH in heavy water, Anal. Chem. **40,** No. 4, 700–706 (Apr. 1968).
Coyle, T. D., Cooper, J., Ritter, J. J., Preparation and some reactions of difluoroborane, Inorg. Chem. **7,** No. 5, 1014–1020 (May 1968).
Coyle, T. D., Ritter, J. J., Reactions of diboron tetrahalides with some unsaturated organometallic compounds, J. Organometallic Chem. **12,** 269–280 (1968).
Crumlish, J. D., Wirth, G. F., A preliminary study of engineering seismology benefits, Environmental Science Services Administration, Coast and Geodetic Survey (Superintendent of Documents, Government Printing Office, Washington, D.C., 20402, Aug. 1967, 35 cents).
Cushen, W. E., Systems engineering—Where did it come from . . . where will it go? Consulting Engr. **30,** No. 3, 130–134 (Mar. 1968).
Cuthill, J. R., McAlister, A. J., Williams, M. L., Soft x-ray spectroscopy of alloys: TiNi and the Ni-Al system, J. Appl. Phys. **39,** No. 5, 2204–2208 (Apr. 1968).
Dehl, R. E., Broadline NMR study of H_2O and D_2O in oriented rayon fibers, J. Chem. Phys. **48,** No. 2, 831–835 (Jan. 15, 1968).
deMain, P. A. D., Marron, B. A., The SOLID SYSTEM. I. A method for organizing and searching files (Proc. Third Annual Colloquium on Information Retrieval, Philadelphia, Pa., May 12–13, 1966), Book, Information Retrieval, A Critical View, Ed. G. Schecter, pp. 243–282 (Thompson Book Co., Washington, D.C., 1967).
DeSimone, D. V., Education for innovation, IEEE Spectrum **5,** No. 1, 83–90 (Jan. 1968).
Deslattes, R. D., Photoionization of the M shell of xenon, Phys. Rev. Letters **20,** No. 10, 483–484 (Mar. 4, 1968).
Deslattes, R. D., X-ray monochromators and resonators from single crystals, Appl. Phys. Letters **12,** No. 4, 133–135 (Feb. 15, 1968).
DeVoe, J. R., Spijkerman, J. J., Mössbauer spectrometry, Anal. Chem. **40,** No. 5, 472R–489R (Apr. 1968).
Dibeler, V. H., Walker, J. A., Mass spectrometric study of the photoionization of small polyatomic molecules (Proc. Conf. on Mass Spectrometry, Berlin, Germany, Sept. 1967), Book, Advances in Mass Spectrometry, **4,** 767–780 (The Institute of Petroleum, London, England, 1968).
Dick, C. E., Marella, A. B., Miller, W. C., A simple beam positioning system for low energy electrons, Nucl. Instr. Methods **60,** 346–348 (1968).
Dickson, P. F., Jones, M. C., Infrared spectral reflectances of metals at low temperatures, Cryogenics **8,** No. 1, 24–29 (Feb. 1968).
Dickson, R. W., Anderson, R. C., Temperature dependence of the elastic moduli of $91Y_2O_3+9ThO_2$ from 25° to 1100 °C, J. Am. Ceram. Soc. **51,** No. 4, 233 (Apr. 1968).
Domsitz, M. G., Transducer activities at NBS (Proc. Fifth Transducer Workshop, National Bureau of Standards, Gaithersburg, Md., Oct. 3–4, 1967), Telemetry Working Group Document 112–68, 11–12 (Secretariat, Range Commanders Council, White Sands Missile Range, New Mexico, 1968).
Durst, R. A., Fluoride microanalysis by linear null-point potentiometry, Anal. Chem. **40,** No. 6, 931–935 (May 1968).
Durst, R. A., May, E. L., Taylor, J. K., Improved technique for the microdetermination of silver by linear null-point potentiometry, Anal. Chem. **40,** No. 6, 977–978 (May 1968).
Eisenhart, C., Gauss, Carl Freidrich, International Encyclopedia of the Social Sciences **6,** 74–81 (The Macmillan Company and The Free Press, New York, N.Y., 1968).
Eisenhauer, C., Gamma radiation fluxes near a ground-air interface using an image source technique, Nucl. Sci. Engr. **32,** 166–177 (1968).
Fatiadi, A. J., Formation of arylazocyclohexane derivatives on acylation of certain inosose phenylhydrazones, Carbohydrate Res. **7,** 89–94 (1968).
Fell, H., Mather, J., Barely faithful algebras, Math. Monthly **72,** No. 9, 1001–1003 (Nov. 1965).
Flynn, D. R., O'Hagan, M. E., Measurements of the thermal conductivity and electrical resistivity of platinum from 100° C to 900 °C. Engelhard Ind. Tech. Bull. **VIII,** No. 4, 117–147 (Mar. 1968).
Foster, B. E., Attenuation of x-rays and gamma rays in concrete, Mater. Res. Std. **8,** No. 3, 19–24 (Mar. 1968).
Fraction, G. F., Tauber, S. J., Walker, J. C., Connection tables from Wiswesser line notation: A partial algorithm, Proc. Wiswesser Line Notation Meeting of the Army Chemical Information and Data Systems Program, Edgewood Arsenal, Watertown, N.J., Oct. 6–7, 1966, Edgewood Arsenal Spec. Publ. EASP–400–8, pp. 139–197 (Jan. 1968).
Franklin, A. D., Born model calculation of defect energies in CaF_2, J. Phys. Chem. Solids **29,** 823–841 (1968).
Goldman, A. J., Nemhauser, G. L., A transport improvement problem transformable to a best-path problem, Transportation Sci. **1,** No. 4, 295–307 (Nov. 1967).
Goldman, D. T., Suggested procedures for utilizing MUFT resonance parameters to include Doppler broadening, Nucl. Sci. Engr. **31,** No. 2, 346–349 (Feb. 1968).
Gray, V. E., Cadoff, B. C., Survey of techniques for evaluating effects of weathering on plastics, Appl. Polymer Symp. **4,** 85–95 (1967).
Halford, D., A general mechanical model for $|f|^\alpha$ spectral density random noise with special reference to flicker noise $1/|f|$, Proc. IEEE **56,** No. 3, 251–258 (Mar. 1968).
Hall, J. L., and Siegel, M. W., Angular dependence of the laser photo-detachment of the negative ions of carbon, oxygen, and hydrogen, J. Chem. Phys. **48,** No. 2, 943–945 (Jan. 15, 1968).
Hall, L. H., Spijkerman, J. J., Lambert, J. L., Preparation and coordination studies of the complex acid, dihydrogen diethylenetriamine-pentaacetatoferrate (III) dihydrate, and several of its metal (I) salts, J. Am. Chem. Soc. **90,** No. 8, 2044–2048 (Apr. 10, 1968).
Hamer, W. J., Battery, electric, Encylopedia Americana **3,** 357–363 (1967).
Hamer, W. J., Using standard cells, Meas. Data **1,** No. 6, 64–67 (Nov.–Dec. 1967).
Hamer, W. J., Wood, R. E., Electrolytic conductivity and electrode processes, Handbook of Physics, Second Ed., Ed. E. U. Condon and H. Odishaw, Part 4. Electricity and Magnetism, ch. **9,** pp. 4–146—4–171 (McGraw-Hill Book Co., New York, N.Y., 1967).
Hammond, H. K., III, Phototube response evaluation, Appl. Opt. **7,** No. 5, 985 (May 1968).
Harvey, M. E., Precision temperature-controlled water bath, Rev. Sci. Instr. **39,** No. 1, 13–18 (Jan. 1968).
Hayward, R. W., Nuclear electromagnetic radiation, Handbook of Physics, Second Ed., Ed. E. U. Condon and H. Odishaw, Part 9. Nuclear Physics, ch. 6, pp. 9–172—9–194 (McGraw-Hill Book Co., New York, N.Y., 1967).
Heydemann, P. L. M., An ultrasonic pressure gage, J. Basic Eng. **89,** No. 3, 551–555 (Sept. 1967).
Hilten, J. S., The National Bureau of Standards interagency transducer project—a progress report (Proc. Fifth Transducer Work-

shop, National Bureau of Standards, Gaithersburg, Md., Oct. 3–4, 1967), Telemetry Working Group Document 112–68, 17–36 (Secretariat, Range Commanders Council, White Sands Missile Range, New Mexico, 1968).

Huget, E. F., Brauer, G. M., Loebenstein, W. V., Apparent heats of wetting and heats of reaction of the components of tooth structure and synthetic fluorapatite, J. Dental Res. **47,** No. 2, 291–301 (Apr. 1968).

Hughes, E. E., A simple technique for the absolute determination of atmospheric oxygen, Environ. Sci. Technol. **2,** No. 3, 201–203 (Mar. 1968).

Hughes, E. E., and Dorko, W. D., Accurate mass spectrometric determination of low concentrations of carbon dioxide in nitrogen, Anal. Chem. **40,** No. 4, 750–755 (Apr. 1968).

Hutchinson, J. M. R., Naas, C. R., Walker, D. H., Mann, W. B., Backscattering of alpha particles from thick metal backings as a function of atomic weight, Intern. J. Appl. Radiation Isotopes **19,** 517–522 (June 1968).

Ives, L. K., Ruff, A. W., Jr., Studies of the effect of annealing on extended dislocation nodes in silver-tin alloys, Phys. Stat. Sol. **27,** 117–123 (1968).

Jensen, M. W., The Fair Packaging and Labelling Act of 1966—its history and its implementation, Monthly Rev. J. Inst. Weights Measures Admin. **76,** No. 4, 3–6 (Apr. 1968).

Johannesen, R. B., Relative signs of NMR coupling constants and isotope shifts in ammonium ions, J. Chem. Phys. **48,** No. 3, 1414–1415 (Feb. 1, 1968).

Judd, D. B., Color vision and colorimetry, Handbook of Physics, Second Ed., Ed. E. U. Condon and H. Odishaw, Part 6. Optics, ch. 4. pp. 6–65—6–77 (McGraw-Hill Book Co., New York, N.Y., 1967).

Kasuya, T., Lafferty, W. J., and Lide, D. R., Jr., Microwave spectrum, structure, boron quadrupole coupling constants, and dipole moment of difluoroborane, J. Chem. Phys. **48,** No. 1, 1–4 (Jan. 1, 1968).

Keller, R. A., Intramolecular energy transfer between triplet states of weakly interacting chromophores. II. Compounds in which the chromophores are perpendicular to each other, J. Am. Chem. Soc. **90,** No. 8, 1940–1944 (Apr. 10, 1968).

Kneissl, G. J., Richmond, J. C., Wiebelt, J. A., A laser source integrating sphere for the measurement of directional hemispherical reflectance at high temperatures (Proc. Thermophysics Specialist Conf. of AIAA, New Orleans, La., Apr. 17–20, 1967), Book, Thermophysics of Spacecraft and Planetary Bodies, Chapt. 1, Radiation Properties of Solids and their Measurement, 20, 177–202 (Academic Press Inc., New York, N.Y., Dec. 1967).

Kokoszka, G. F., Brinckman, F. E., Electron paramagnetic resonance studies of phosphorus-containing reactive intermediates, Chem. Commun. **7,** 349–350 (Apr. 10, 1968).

Kuriyama, M., Dynamical calculation of the x-ray diffraction profile from a distorted crystal, J. Appl. Phys. **39,** No. 4, 2162–2163 (Mar. 1968).

Kushner, L. M., The role of the Institute for Applied Technology (Proc. Fifth Transducer Workshop, National Bureau of Standards, Gaithersburg, Md., Oct. 3–4, 1967), Telemetry Working Group Document 112–68, 5–10 (Secretariat, Range Commanders Council, White Sands Missile Range, New Mexico, 1968).

Larsen, N. T., 50 microdegree temperature controller, Rev. Sci. Instr. **39,** No. 1, 1–12 (Jan. 1968).

Lias, S. G., and Ausloo, P., Gas-phase photolysis and radiolysis of isobutane, J. Chem. Phys. **48,** No. 1, 392–400 (Jan. 1, 1968).

Little, J. L., Mooers, C. N., Standards for user procedures and data formats in automated information systems and networks, AFIPS Conf. Proc. Spring Joint Computer Conf., Atlantic City, N.J., Apr. 30–May 2, 1968, 32, 89–94 (Thompson Book Co., Washington, D.C., 1968).

Lundgren, F. A., Nargolwalla, S. S., Use of a dual sample-biaxial rotating assembly with a pneumatic tube transfer system for high precision 14-MeV neutron activation analysis, Anal. Chem. **40,** No. 4, 672–677 (Apr. 1968).

McKinley, J. D., Effects of the reaction with bromine on nickel surface morphology, Surface Sci. **10,** No. 2, 287–290 (May 1968).

Mahler, R. J., James, L. W., Nuclear electric hexadecapole phonon interaction (Proc. XIV Colloque Ampere, Sept. 6–11, 1966, Ljubljana, Yugoslavia), Book, Magnetic Resonance and Relaxation, Ed. R. Blinc, pp. 938–945 (North-Holland Publ. Co., Amsterdam, The Netherlands, 1967).

Maienthal, E. J., and Taylor, J. K., Polarographic methods in determination of trace inorganics in water, Book, Trace Inorganics In Water, ch. 10, Advances in Chemistry Series **73,** 172–182 (American Chemical Society, Washington, D.C., Apr. 1968).

Maki, A. G., Assignment of some DCN and HCN laser lines, Appl. Phys. Letters **12,** No. 4, 122–124 (Feb. 15, 1968).

Maki, A. G., Thibault, R., Analysis of some perturbations in the ν_4 band of methyl chloride, J. Chem. Phys. **48,** No. 5, 2163–2167 (Mar. 1, 1968).

Manning, J. R., Correlation factors for diffusion of vacancies in binary alloys (Proc. Subconference on Lattice Defects, Honolulu, Hawaii, Sept. 1965), Book, Lattice Defects and Their Interactions, Ed. R. R. Hasiguti, pp. 269–289 (Gordon and Breach, New York, N.Y., 1967).

Margoshes, M., Scribner, B. F., Emission spectrometry, Anal. Chem. **40,** No. 5, 223R–246R (Apr. 1968).

Marron, B. A., deMaine, P. A. D., Automatic Data Compression, Commun. ACM 10, No. 11, 711–715 (Nov. 1967).

Marvin, R. S., Twenty years of rheology, Phys. Today (20th Anniversary Edition) **21,** No. 5, 52–53 (May 1968).

Mattis, R. L., Phillips, W. E., Bullis, W. M., Measurement and interpretation of carrier lifetime in silicon and germanium, Tech. Rept. AFML–TR–68–81 (Air Force Materials Laboratory, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, July 1968).

Melmed, A. J., Field emission shadow microscopy, Appl. Phys. Letters **12,** No. 3, 100–102 (Feb. 1, 1968).

Menke, J. L., Schrack, R. A., Feeble field feeler, Nucl. Instr. Methods **57,** 158–159 (1967).

Mies, F. H., Quantum oscillations in the shape of pressure-broadened atomic lines, J. Chem. Phys. **48,** No. 1, 482–494 (Jan. 1, 1968).

Miller, K. J., and Krauss, M., Born inelastic differential cross sections in H_2, J. Chem. Phys. **47,** No. 10, 3754–3762 (Nov. 15, 1967).

Milligan, D. E., Jacox, M. E., Matrix-isolation study of the reaction of atomic and molecular fluorine with carbon atoms. The infrared spectra of normal and ^{13}C-substituted CF_2 and CF_3, J. Chem. Phys. **48,** No. 5, 2265–2271 (Mar. 1, 1968).

Moore-Sitterly, C. E., Commission 14—Fundamental spectroscopic data report of meetings, 23 and 25 August 1967, Trans. Intern. Astron. Union 13B, 87–91 (1968).

Motz, J. W., and Sparrow, J. H., A simple device for the energy and current measurement of an accelerator electron beam, Record of the IEEE 9th Ann. Symp. Electron, Ion, and Laser Beam Technology, Berkeley, Calif., May 9–11, 1967, pp. 34–41 (San Francisco Press Inc., San Francisco, Calif., May 1967).

Nargolwalla, S. S., Crambes, M. R., DeVoe, J. R., A technique for the evaluation of systematic errors in the activation analysis for oxygen with 14-MeV neutrons, Anal. Chem. **40,** No. 4, 666–671 (Apr. 1968).

Oppenheim, I., Shuler, K. E., Weiss, G. H., On the decay of initial correlations in stochastic processes, J. Chem. Phys. **46,** No. 10, 4100–4111 (May 15, 1967).

Page, C. H., Electronic circuits, Handbook of Physics, Second Ed., Ed. E. U. Condon and H. Odishaw, Part 4. Electricity and Magnetism, Ch. 4, pp. 4–45—4–52 (McGraw-Hill Book Co., New York, N.Y., 1967).

Pichanick, F. M. J., Simpson, J. A., Resonances in the total cross sections for metastable excitation of noble gases by electron impact, Phys. Rev. **168,** No. 1, 64–70 (Apr. 5, 1968).

Radford, H. E., Evenson, K. M., Paramagnetic resonance spectrum of metastable (^2D) atomic nitrogen, Phys. Rev. **168,** No. 1, 70–74 (Apr. 5, 1968).

Rebbert, R. E., and Ausloos, P., Photolysis of methyl iodide in matrices of organic compounds at 20° and 77 °K. Reactions of hot methyl radicals, J. Chem. Phys. **48,** No. 1, 306–311 (Jan. 1, 1968).

Reimann, C. W., Zocchi, M., The structure of the trinuclear cation *bis*-[μ-(*tri*-1,2,4-triazolo-N^1,N^2)-triquonickel]nickel, Chem. Commun. **5,** 272 (Mar. 6, 1968).

Rosenstock, H. M., Theory of mass spectra—A general review (Proc. Conf. on Mass Spectrometry, Berlin, Germany, Sept. 1967), Book, Advances in Mass Spectrometry 4, 523–545 (The Institute of Petroleum, London, England, 1968).

Rush, J. J., Flowtow, H. E., Vibration spectra of vanadium hydride in three crystal phases by inelastic neutron scattering, J. Chem. Phys. **48,** No. 8, 3795–3804 (Apr. 15, 1968).

Saunders, J. B., Precision measurements, Book, Advanced Optical Techniques, ch. 1, pp. 3–22 (North-Holland Publ. Co., Amsterdam, The Netherlands, 1967).

Schneider, W. E., Garvey, J. A., Procedures for determining the spectral responsivity of an infrared radiometer, Appl. Opt. **7,** No. 6, 1141–1148 (June 1968).

Sengers, J. M. H. L., On the determination of virial coefficients from PVT data, Proc. 4th Symp. Thermophysical Properties, University of Maryland, College Park, Md., Apr. 1–4, 1968, pp. 37–44 (American Society of Mechanical Engineers, New York, N.Y., 1968).

Sengers, J. M. H. L., Vicentini-Missoni, M., Thermodynamic anomalies near the gas-liquid critical point, Proc. 4th Symp. Thermophysical Properties, University of Maryland, College Park, Md., Apr. 1–4, 1968, pp. 79–86 (American Society of Mechanical Engineers, New York, N.Y., 1968).

Sengers, J. V., Sengers, J. M. H. L., The critical region, Chem. Eng. News **46,** No. 25, 104–118 (June 10, 1968).

Shuler, K. E., Microscopic and macroscopic approach to chemical rate processes (Proc. 7th AGARD Colloquium, Oslo, Norway, May 16–20, 1966), Book, Recent Advances in Aerothermochemistry, Ed. I. Glassman, 1, No. 12, 1–21 (Technical Editing and Reproduction Ltd., London, England, 1967).

Smit, J., Ogburn, F., Bechtoldt, C. J., Multiple twin structures in electrodeposited silver dendrites, J. Electrochem. Soc. **115,** No. 4, 372–374 (Apr. 1968).

Spijkerman, J. J., Hall, H. L., Lambert, J. L., Coordination studies by Mössbauer spectroscopy of some metal (I) salts of the complex acid, hydrogen aquoethylenediaminetetraacetatoferrate (III), J. Am. Chem. Soc. **90,** No. 8, 2039–2043 (Apr. 10, 1968).

Stern, J., Activities of the Basic Instrumentation Section (Proc. Fifth Transducer Workshop, National Bureau of Standards, Gaithersburg, Md., Oct. 3–4, 1967), Telemetry Working Group Document 112–68, 13–15 (Secretariat, Range Commanders Council, White Sands Missile Range, New Mexico, 1968).

Stern, K. H., Ion diffusion into fused silica from molten salts, J. Phys. Chem. **72,** No. 6, 2256–2259 (June 1968).

Stern, K. H., Membrane potentials and ion selectivity of fused silica in molten salts, J. Phys. Chem. **72,** No. 6, 1963–1975 (June 1968).

Swanson, N., Powell, C. J., Excitation of L-shell electrons in Al and Al_2O_3 by 20-keV electrons, Phys. Rev. **167,** No. 3, 592–600 (Mar. 15, 1968).

Swartzendruber, L. J., Bennett, L. H., Hyperfine field distributions in $(ZrCo_2)$ $(ZrFe_2)_{1-}$ alloys, J. Appl. Phys. **39,** No. 2, Part II, 1323–1324 (Feb. 1, 1968).

Swartzendruber, L. J., Bennett, L. H., Line profiles in the nuclear magnetic resonance and Mössbauer effect of $TiFe_{1-x}Co_x$ alloys, J. Appl. Phys. **39,** No. 5, 2215–2220 (Apr. 1968).

Swartzendruber, L. J., Bennett, L. H., Mössbauer line structure in "non-magnetic" Cu-Fe and Cu-Ni-Fe alloys, Physics Letters **27A,** 141–142 (1968).

Swindells, J. F., Thermometry and pyrometry, Handbook of Physics, Second Ed., Ed. E. U. Condon and H. Odishaw, Part 5, Heat and Thermodynamics, ch. 3, pp. 5–34—5–44 (McGraw-Hill Book Co., New York, N.Y., 1967).

Sze, W. C., Comparator for calibration of inductive voltage dividers from 1 to 10 kHz, ISA Trans. **6,** No. 4, 263–267 (Oct. 1967).

Tauke, J., Litovitz, T. A., and Macedo, P. B., Viscous relaxation and non-Arrhenius behavior in B_2O_3, J. Am. Ceramic Soc. **51,** No. 3, 158–163 (Mar. 1968).

Tighe, N. J., Hyman, A., Transmission electron microscopy of alumina ceramics (Proc. Intern. Symp. Anisotropy in Single-Crystal Refractory Compounds, Dayton, Ohio, June 13–15, 1967), Book, Anisotropy in Single-Crystal Refractory Compounds, Ed. F. W. Vahldiek and S. A. Mersol, 2, 121–136 (Plenum Press Inc., New York, N.Y., 1968).

van Reuth, E. C., and Waterstrat, R. M., Atomic ordering in binary A15-type phases, Acta Cryst. **B24,** Part 2, 186–196 (Feb. 1968).

Watson, R. E., Bennett, L. H., Freeman, A. J., Origin of solvent Knight shifts in alloys, Phys. Rev. Letters **20,** No. 13, 653–656 (Mar. 25, 1968).

Wiederhorn, S. M., Shorb, A. M., Moses, R. L., Critical analysis of the theory of the double cantilever method of measuring fracture-surface energies, J. Appl. Phys. **39,** No. 3, 1569–1572 (Feb. 15, 1968).

Woelfel, J. B., and Paffenbarger, G. C., Evaluation of complete dentures lined with resilient silicone rubber, J. Am. Dent. Assoc. **76,** 582–590 (Mar. 1968).

Yakowitz, H., Evaluation of specimen preparation and the use of standards in electron probe microanalysis, Am. Soc. Testing Mater. Spec. Tech. Publ. **430,** Fifty Years of Progress in Metallographic Techniques, 383–408 (1968).

Youden, W. J., Statistical design of experiments, Handbook of Physics, Second Ed., Ed. E. U. Condon and H. Odishaw, Part 1. Mathematics, ch. 13, pp. 1–198—1–201 (McGraw-Hill Book Co., New York, N.Y., 1967).

**Publications for which a price is indicated are available by purchase from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (foreign postage, one-fourth additional). The NBS non-periodical series are also available from the Clearinghouse for Federal Scientific and Technical Information, Springfield, Va. 22151. Reprints from outside journals and the NBS Journal of Research may often be obtained directly from the authors.*

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 72B, No. 4, October–December 1968

INDEX TO VOLUME 72B

January to December 1968

A

Aissen, M., Means and the minimization of errors 1
Algebraic equations 23
Allocating service periods to minimize delay time 215
Allocation 215
Alt. F. L., Kirsch, R. A., Citation searching and bibliographic coupling with remote on-line computer access 61
Analysis of variance 33
Arithmetic mean 1
Atchison, T. A., Gray, H. L., A note on the *G*-transformation 29

B

BBGKY hierarchy equations 127
Bessel functions 61
Bibliographies 145
Bipartite graphs 79
Born approximation 33

C

Calibration 7
Cancellation 201
Cauchy inequalities 61
Characteristics 61
Citation searching and bibliographic coupling with remote on-line computer access 145
Citations 239
Class of thickness-minimal graphs 239
Commutator subgroup 159
Comparison of finite difference computations of natural convection 33
Computer 79
Conjectures 239
Connectivity 159
Contingency tables 33
Correlation coefficient 79
Coulomb field 239
Cut-set 33

D

Degrees of freedom 43
Demand 103
Differentiable generalized inverse 109
Differentiable generalized inverses 135
Diffusion 23
Diophantine equation $P(x, y)=(xy+d)z$ 23
Diophantine equations 33
Distribution of the sample correlation coefficient with one variable fixed 37
Dirac and Klein-Gordon equations 33
Distribution 115

E

Eigenvalues 43
Elasticity 93
Enclosures 159
Equilibrium 159
Estimation of cell frequencies from marginals 109
Evans, J. W., Hearon, J. Z., Differentiable generalized inverses 103
—, —, On spaces and maps of generalized inverses 103

F

Finite-difference 201
Finite-difference equations 33
Fixed variable 201
Fluid flow 229
Fong, J. T., Smith, J. C., On the coupling of longitudinal and transverse waves in a linear three-element viscoelastic string subjected to transverse impact 159
103,
109

G

Gelman, H., The second orthogonality conditions in the theory of proper and improper rotations. I. Derivation of the conditions and of their main consequences 1
Generalized independence 93
Generalized inverse 43
Generalized inverses 145, 239
Generalized inverses and solutions of linear systems 29
Geometric mean 239
Goldberg, K., Goldman, A. J., Newman, M., The probability of an equilibrium point 145
Goldman, A. J., Minimax error selection of a discrete univariate distribution with prescribed componentwise ranking 61
—, *Goldberg, K., Newman, M.*, The probability of an equilibrium point 61
—, *McLynn, J. M., Meyers, P. R., Watkins, R. H.*, Analysis of a market split model 61
—, *Meyers, P. R.*, Minimax error selection of a discrete univariate distribution with prescribed componentwise bounds 61
Graph theory 61
Gray, H. L., Atchison, T. A., A note on the *G*-transformation 61
Grossman, J. W., Hobbs, A. M., A class of thickness-minimal graphs 61
—, —, Thickness and connectivity in graphs 61

H

Harmonic mean 61
Hearon, J. Z., Generalized inverses and solutions of linear systems 61
—, *Evans, J. W.*, Differentiable generalized inverses 61
—, —, On spaces and maps of generalized inverses 61
Hermitian matrices 61
Hobbs, A. M., Grossman, J. W., A class of thickness-minimal graphs 61
—, —, Thickness and connectivity in graphs 61
Hogben, D., The distribution of the sample correlation coefficient with one variable fixed 61
Horn, W. A., Allocating service periods to minimize delay time 61
Hypothesis testing 61

I

Improper integrals 61
Information retrieval 61
Information theory 61

Page

Integers ... 23
Interaction ... 159
Interaction in multidimensional contingency tables: an information theoretical approach ... 159
Interlacing theorems ... 7
Ion flow tubes ... 135
Iterated integrals ... 5

J

Jarvis, S., Jr., On the diffusion of an ion sheet in Poiseuille flow ... 135

K

Kirsch, R. A., Alt, F. L., Citation searching and bibliographic coupling with remote on-line computer access ... 61
Ku, H. H., Kulback, S., Interaction in multidimensional contingency tables: an information theoretical approach ... 159
Kullback, S., Ku, H. H., Interaction in multidimensional contingency tables: an information theoretical approach ... 159

L

Lehner, J., On the multipliers, of the Dedekind modular function ...
Linear algebra ... 103
Linear programs ...
Linear systems ...
Linear viscoelasticity ... 201

M

Mathematical basis for the plasma kinetic equations (BBGKY) ... 127
Mathematical economics ... 43
Mathematical models ...
Mathematics ... 145, 239
Matrices ... 7, 229
Matrix ... 103, 109, 115,
Matrix inequalities ... 7
Matrix theory ... 7
Maximon, L. C., Simple analytic expressions for the total Born approximation cross section for pair production in a Coulomb field ... 79
McLynn, J. M., Goldman, A. J., Meyers, P. R., Watkins, R. H., Analysis of a market split model ... 43
Means ... 1
Means and minimization of errors ... 1
Meyers, P. R., Goldman, A. J., Minimax error selection of a discrete univariate distribution with prescribed componentwise bounds ...
—, *McLynn, J. M., Goldman, A. J., Watkins, R. H.*, Analysis of a market split model ... 43
Minimality ... 239
Minimax error selection of discrete univariate distribution with prescribed componentwise bounds ...
Minimax error selection of a discrete univariate distribution with prescribed componentwise ranking ...
Minimax estimation ...
Moderate energy cross section ... 79
Modular group ...
Multiplier ...
Myers, V. W., Solutions of the time-dependent Klein-Gordon and Dirac equations for a uniform electric field ... 37

N

Natural convection ...
Newman, M., Goldberg, K., Goldman, A. J., The probability of an equilibrium point ... 93
Noncentral veta variable ... 33
Noncentrality ... 33
Nonlinear transformation ... 29
Note on the *G*-transformation ... 29

Page

ν-Derivatives ...
Numerical ...

O

On spaces and maps of generalized inverses ... 103
On the coupling of longitudinal and transverse waves in a linear three-element viscoelastic string subjected to transverse impact ... 201
On the diffusion of an ion sheet in Poiseuille flow ... 135
On the multipliers of the Dedekind modular function ...
On the signs of the ν-derivatives of the modified Bessel function $I_\nu(x)$ and $K_\nu(x)$ ...
Operations research ...
Orthogonal transformations ... 229
Osgood, C. F., The diophantine equation $P(x, y) - (xy - d)z$ ... 23
Pair production ... 79
Partial differential equations ... 43, 201
Planar graphs ... 145
Principal submatrices ... 7, 115,
Principal submatrices III: Linear inequalities ... 7
Principal submatrices V: Some results concerning principal submatrices of arbitrary matrices ... 115
Principal submatrices VII: Further results concerning matrices with equal principal minors ...
Probabilistic basis for kinetic equations ... 127
Probability ... 93
Probability distribution ...
Probability of an equilibrium point ... 93

Q

Q variate ... 33
Queueing theory ... 215

R

Racah cross section ... 79
Rank ... 115,
Relative error ... 1
Remainders ... 5
Remote consoles ... 61
Reudink, D. O., On the signs of the ν-derivatives of the modified Bessel functions $I_\nu(x)$ and $K_\nu(x)$ ...
Rotation ... 229

S

Scheduling ... 215
Second-order interaction ... 159
Second orthogonality conditions in the theory of proper and improper rotations. I. Derivation of the conditions and of their main consequences ... 229
Shisha, O., On Taylor's theorem ... 5
Simple analytical expressions for the total Born approximation cross section for pair production in a Coulomb field ... 79
Smith, J. C., Fong, J. T., On the coupling of longitudinal and transverse waves in a linear three-element viscoelastic string subjected to transverse impact ... 201
Solutions of the time-dependent Klein-Gordon and Dirac equations for a uniform electric field ... 37
Sopka, J. J., Mathematical basis for the plasma kinetic equations (BBGKY) ... 127
Switching theory ... 215
Symmetric matrix ... 115,

T

Taylor series ... 5
Taylor's theorem ... 5
Theory of games ... 93
Thickness and connectivity in graphs ... 239
Thickness of graphs ... 145, 239
Thompson, R. C., Principal submatrices III: Linear inequalities ... 7

Page

—, Principal submatrices V: Some results concerning principal submatrices of arbitrary matrices 115
—, Principal submatrices VII: Further results concerning matrices with equal principal minors .....
Three variables 23
t-Minimal graphs 145
Time-dependent solutions 37
Timesharing 61
Torrance, K. E., Comparison of finite-difference computations of natural convection .....
Total photon cross section 79
Traffic flow 215
Transportation theory 215
Transverse impact 201

Page

U

Uniform electric field 37

W

Watkins, R. H., McLynn, J. M., Goldman, A. J., Meyers, P. R., Analysis of a market split model 43
Wave propagation 201
Waves in strings 201
Word .....

U.S. GOVERNMENT PRINTING OFFICE : 1969 OL—321-483

A.B.C.

UNITED STATES
GOVERNMENT PRINTING OFFICE
DIVISION OF PUBLIC DOCUMENTS
WASHINGTON, D.C. 20402

OFFICIAL BUSINESS

POSTAGE AND FEES PAID
U.S. GOVERNMENT PRINTING OFFICE

JOURNAL OF RESEARCH of the National Bureau of Standards Vol. 72B, No. 4 October–December 1968
B. Mathematical Sciences

Contents

	Page
Principal submatrices VII: Further results concerning matrices with equal principal minors. R. C. Thompson.	249
On the multipliers of the Dedekind modular function. Joseph Lehner.	253
Minimax error selection of a discrete univariate distribution with prescribed componentwise bounds. A. J. Goldman and P. R. Meyers.	263
Minimax error selection of a discrete univariate distribution with prescribed componentwise ranking. A. J. Goldman.	273
On the signs of the ν-derivatives of the modified Bessel functions $I_\nu(x)$ and $K_\nu(x)$. D. O. Reudink.	279
Comparison of finite-difference computations of natural convection. K. E. Torrance.	281
Generalized inverses and solutions of linear systems. John Z. Hearon.	303
Publications of the National Bureau of Standards.	309
Index to volume 72B, January to December 1968.	315

Library of Congress Catalog Card Number: 62–4414

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Price 75c (single copy). Subscription price: $2.25 a year; 50c additional for foreign mailing.

UNITED STATES GOVERNMENT PRINTING OFFICE, WASHINGTON : 1968